Volume 5, No. 3 | May–June 2022

# RAIL

# The Journal of Robotics, Artificial Intelligence & Law

fastcase FULL COURT PRESS

RAIL
The Journal of Robotics, Artificial Intelligence & Law
Volume 5, No. 3 | May–June 2022

**EDITOR-IN-CHIEF**

**Steven A. Meyerowitz**
*President, Meyerowitz Communications Inc.*

**EDITOR**

**Victoria Prussen Spears**
*Senior Vice President, Meyerowitz Communications Inc.*

**BOARD OF EDITORS**

**Miranda Cole**
*Partner, Covington & Burling LLP*

**Kathryn DeBord**
*Partner & Chief Innovation Officer, Bryan Cave LLP*

**Melody Drummond Hansen**
*Partner, O'Melveny & Myers LLP*

**Paul B. Keller**
*Partner, Allen & Overy LLP*

**Garry G. Mathiason**
*Shareholder, Littler Mendelson P.C.*

**Elaine D. Solomon**
*Partner, Blank Rome LLP*

**Linda J. Thayer**
*Partner, Finnegan, Henderson, Farabow, Garrett & Dunner LLP*

**Edward J. Walters**
*Chief Executive Officer, Fastcase Inc.*

**John Frank Weaver**
*Attorney, McLane Middleton, Professional Association*

THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW (ISSN 2575-5633 (print) /ISSN 2575-5617 (online) at $495.00 annually is published six times per year by Full Court Press, a Fastcase, Inc., imprint. Copyright 2022 Fastcase, Inc. No part of this journal may be reproduced in any form—by microfilm, xerography, or otherwise—or incorporated into any information retrieval system without the written permission of the copyright owner. For customer support, please contact Fastcase, Inc., 711 D St. NW, Suite 200, Washington, D.C. 20004, 202.999.4777 (phone), 202.521.3462 (fax), or email customer service at support@fastcase.com.

Publishing Staff
Publisher: Morgan Morrissette Wright
Production Editor: Sharon D. Ray
Cover Art Design: Juan Bustamante

Cite this publication as:

The Journal of Robotics, Artificial Intelligence & Law (Fastcase)

This publication is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If legal advice or other expert assistance is required, the services of a competent professional should be sought.

Copyright © 2022 Full Court Press, an imprint of Fastcase, Inc.

All Rights Reserved.

A Full Court Press, Fastcase, Inc., Publication

Editorial Office

711 D St. NW, Suite 200, Washington, D.C. 20004
https://www.fastcase.com/

POSTMASTER: Send address changes to THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW, 711 D St. NW, Suite 200, Washington, D.C. 20004.

## Articles and Submissions

Direct editorial inquiries and send material for publication to:

Steven A. Meyerowitz, Editor-in-Chief, Meyerowitz Communications Inc.,
26910 Grand Central Parkway, #18R, Floral Park, NY 11005, smeyerowitz@
meyerowitzcommunications.com, 631.291.5541.

Material for publication is welcomed—articles, decisions, or other items of interest to attorneys and law firms, in-house counsel, corporate compliance officers, government agencies and their counsel, senior business executives, scientists, engineers, and anyone interested in the law governing artificial intelligence and robotics. This publication is designed to be accurate and authoritative, but neither the publisher nor the authors are rendering legal, accounting, or other professional services in this publication. If legal or other expert advice is desired, retain the services of an appropriate professional. The articles and columns reflect only the present considerations and views of the authors and do not necessarily reflect those of the firms or organizations with which they are affiliated, any of the former or present clients of the authors or their firms or organizations, or the editors or publisher.

QUESTIONS ABOUT THIS PUBLICATION?

For questions about the Editorial Content appearing in these volumes or reprint permission, please contact:

Morgan Morrissette Wright, Publisher, Full Court Press at mwright@fastcase.com
or at 202.999.4878

For questions or Sales and Customer Service:

Customer Service
Available 8 a.m.–8 p.m. Eastern Time
866.773.2782 (phone)
support@fastcase.com (email)

Sales
202.999.4777 (phone)
sales@fastcase.com (email)
ISSN 2575-5633 (print)
ISSN 2575-5617 (online)

# The Legal Implications of Explaining Artificial Intelligence

David van Boven, Paul B. Keller, Harriet Ravenscroft, Jill Ge, Wentao Zhai, and Arwen Zhang*

*This article reviews current legal requirements for explaining artificial intelligence and its legal implications and provides an analysis of legislative developments in the European Union, United States, and China.*

HAL 9000: "I'm sorry, Dave. I'm afraid I can't do that."

—*2001: A Space Odyssey*

## Artificial Intelligence

The concept of artificial intelligence ("AI") has captured our imagination for generations in books and movies. The underlying technology has been with us for some time, but now society is on the precipice of having this technology in widespread use throughout our daily lives. From healthcare to finance to transportation to farming, AI now touches every aspect of our lives, and that use promises only to increase and become more and more sophisticated.

The promise of AI technology is plain—more efficient use of limited resources to achieve results that are as good, or better, than a human could achieve, and to do so in a fraction of the time. Trust in this technology, however, is still very much in development. As has been repeatedly demonstrated, society is cautious when it comes to new technologies, and it frequently takes years (if not generations) to adopt new technology, even when the facts so plainly indicate that it is better and safer. AI is proving to be no exception.

Some of the issues of trust are rooted in the fundamental nature of the technology. Many of the deep learning neural networks utilize algorithms for machine learning that are unable to be examined after a decision/prediction has been made. These networks

*Robotics, Artificial Intelligence & Law* / May–June 2022, Vol. 5, No. 3, pp. 157–196.
© 2022 Full Court Press. All rights reserved.
ISSN 2575-5633 (print) /ISSN 2575-5617 (online)

rearrange their connections and the strength of those connections in response to patterns they see in the data they are processing, which means that once a neural network has been trained, even the designer of that network cannot know exactly how it is doing what it does. People need the power to disagree with, or reject, an automated decision, but this cannot be done if the user is unable to understand an AI system's decision. This lack of transparency therefore creates issues of trust for the user, and is commonly referred to as the "black box" problem.

This issue is not limited to any one particular application of AI, but rather touches on the full spectrum of potential uses. As autonomous vehicles ("AVs") are expected to roam the streets and smart cameras are expected to be increasingly more present in cities and towns, the understanding of what happens under the hood may fade away. As the deployment of AI accelerates, human understanding of AI, "and also the ability to give informed consent, could be left behind."[1] To be able to more fully rely on this technology, gaining trust may require greater transparency and the accuracy of AI needs to improve. As has been previously argued, there is a trade-off between accuracy and explainability.

## Explainable AI

This article addresses one of the myriad ways that the industry is trying to address this problem: Explainable AI ("XAI"). XAI seeks to allow humans to understand why the system did what it did and not something else. Akin to an audit trail but much more complex, XAI may offer a means for humans to "check" how the AI is operating and identify the fault (and potential liability) if something goes wrong. The use—and lack thereof—of this technology, therefore, has a number of legal implications. Set out below is a thorough overview of the technology and a discussion of (1) the legal requirements and (2) the legal implications in three of the major markets of the world—the United States, Europe, and China.

AI is here to stay and disputes relevant to clients are inevitable. For instance, in the Dutch courts Uber drivers requested an insight into the company's algorithms. It seems likely that regulators and courts will want to know what is going on "inside" a piece of AI. A company might choose to ignore the risk of having to (i.e., being legally obligated to) provide this information, but the penalties (reputational with respect to ethical issues, and financial/operational with respect to legal issues) might be severe. Even if a client does have an XAI system, what might still go wrong?

Many have written about the negative effects of AI. Research has shown that "image search results for certain jobs exaggerate gender stereotypes."[2] Zuiderveen Borgesius argued that while AI may have discriminatory effects—for instance, in the case of search engine image results—the AI itself is not inherently evil.[3] Rather, the dataset may contain biases caused by human biases. AI is merely a complex system, not a nondeterminative system. Any bias in output can therefore only reflect bias in input data or in how the AI is trained. Similarly, an AV is autonomous only in the sense that the end user has to make fewer decisions. Responsibility is not transferred to the AI, however, but to the developer of the AI; in a sense, all future decisions of the AV about how to act in certain situations have been made by the developer, via coding and the choice of training set, before the AV leaves the factory.

Technology brings ethical dilemmas. Nyholm and Smids posed the following scenario:[4] a self-driving car with five passengers approaches another car swerving out of its lane and heading toward the self-driving car. The self-driving car senses the trajectory of the oncoming car. It calculates that a collision is inevitable and will kill the five passengers, unless the self-driving car turns sharply toward the pavement, where a pedestrian is walking. The pedestrian will die from the impact. In this scenario, the humans in the self-driving car cannot take control of the car and thus the car's AI will need to decide. This kind of scenario concerning crashing algorithms,[5] in the literature called an "applied trolley problem"[6] or "collision management,"[7] offers a stark example of the ethical dilemmas that an AI may be faced with.[8]

However, Roff argues that thinking about the trolley problem distracts from understanding the processes of the AI.[9] A computer program is incapable of "having morals" independent of the opinions/biases put into it. As Goodman and Flaxman pointed out, several studies have focused on algorithmic profiling: by explaining the AI, it is possible to both identify and implement interventions to correct for discrimination.[10] By explaining the AI, users could better understand and trust the machine learning capability.[11] But explainability is traded off with accuracy, some argue. It is accepted that explainable models/basic machine learning algorithms, such as decisions, are easily understandable by the human brain, we can simply follow the path the decision tree made to reach the decision.[12] However, these models are not as accurate as they are more simplistic. When we utilize deep learning neural networks, model accuracy increases, but as these models are far more complex, their explainability decreases. Here is the trade-off that has to be made: Do we sacrifice explainability to produce a more accurate model?

It will be argued below that data protection law is not conclusive on the legal requirement of explaining AI. Rather, similar to what Hacker et al. argued,[13] it is expected that XAI will become a legal requirement in other legal domains, such as from a civil liability perspective. Furthermore, sector-specific regulations, such as the draft Digital Services Act, will include transparency requirements regarding the underlying parameters of advertising. Furthermore, the draft AI Regulation contains human oversight, transparency, and traceability requirements for high-risk AI applications.

## Data Protection

The EU's data protection law, the General Data Protection Regulation ("GDPR"), regulates the processing of personal data.[14] According to the European Data Protection Board ("EDPB"): "any processing of personal data through an algorithm falls within the scope of the GDPR."[15] If an AI system processes personal data, the GDPR may apply, but not all AI systems process personal data. Still, even if an AI system is not designed to process personal data, the line between personal data and non-personal data is increasingly unclear, which may be caused by a lack of complete and permanent anonymization and re-identification risks from aggregated datasets.[16]

One can take the collision management scenario one step further and consider what personal data is needed to make decisions on a basis more sophisticated than a simple utilitarian one (i.e., kill the fewest people), in which case access to personal data is likely needed: age (are younger people of more value than older people?), family life (are parents of more value than people with no dependants?), and health (is someone in good health of more value than someone with a preexisting condition?). The collision management thought experiment highlights the importance of personal data. Modern day cars are equipped with cameras, global positioning systems and communication capabilities,[17] while autonomous vehicles are packed with even more sensors.[18]

The GDPR sets out rights for data subjects that can invoke such rights vis-à-vis the processors of their personal data. Articles 13-15 provide rights to "meaningful information about the logic involved" in the case of automated decision-making. These articles are referred to by some as the right to an explanation.[19] More specifically, Article 22 of the GDPR provides that automated decision-making that significantly affects the data subject is prohibited unless the decision-making is contractually required and if the data subject rights are properly safeguarded, or when the data subject has consented to it. The functionality and effectiveness of these rights are highly debated, with some arguing that these rights are too limited and too unclear to be meaningful.[20]

Recital 71 points to the existence of an individual right to an "explanation" of AI decisions.[21] But recital 63 limits such right: "that right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software." Some have argued that recitals, even though they provide an explanation to the substantive articles, do not create legal effect, so a right cannot be deduced solely from the recitals.[22] Nevertheless, the predecessor of the EDPB, the working party group 29, stated in its guidelines that such reasoning should be more nuanced: a data subject may have a right to "obtain an explanation of the decision reached after such assessment and to challenge the decision."[23] The EDPB as such indicates the need for XAI.

The existence of automated decision-making is hard to prove, however. In a Dutch court case in March 2021, it was ruled that Uber's termination of driver agreements in a case of fraud detection did not constitute "automatic decision making" because the termination was only temporary, and therefore did not significantly affect the data subjects.[24] The Uber case shows that it is hard to argue in a specific case whether the automated decision applies in the first place, to then trigger the right to meaningful information.

European privacy law is currently uncertain when it comes to XAI.[25] Some transparency is required, however. The Court of the Hague ruled in February 2020 that the System Risk Indication ("SyRI"), an AI tool used by the Dutch government to combat fraud in benefits claims, did not comply with Article 8 of the European Convention on Human Rights (i.e., the right to respect private life) and was disproportionate to the aims it sought to achieve because of the lack of transparency of the AI.[26] A number of civil society organizations, including the Dutch Lawyers' Committee for Human Rights, and two citizens brought these proceedings against the Dutch State. The court ruled that analysis of data using new technologies can have a profound impact on the private lives of those to whom the data relates. The legislator therefore also bears a special responsibility in the case of the deployment of an instrument such as SyRI: it is difficult for a data subject to oversee the exact impact of the instrument on his or her private life.[27] In the opinion of the district court, the principle of transparency was insufficiently respected in the SyRI case. The district court found that the SyRI application did not provide any information whatsoever about the factual data that may make the presence of a certain circumstance plausible, that is, which objective factual data may justifiably lead to the conclusion that there is an increased risk.[28] The SyRI case shows that, going beyond the realm of automatic decision-making, a complete lack of transparency runs the risk of a ruling declaring the AI system unlawful.

There is however, no clear definition of what XAI means in terms of the GDPR. The GDPR's "right to meaningful information about the logic of processing" is in any event inconclusive when it comes to a clear legal requirement for XAI. As Hacker et al. argued similarly, it is potentially more relevant to expect legal requirements arising in other areas of law.[29]

Unlike the European Union, the United States currently does not have one comprehensive law at the federal level that governs data protection across industries. Instead, there are a number of different, sector-specific laws at the federal level, largely in the healthcare and financial industries. These laws include the U.S. Privacy Act of 1974 ("USPA," personal data held by the government),[30] the Health Insurance Portability and Accountability Act ("HIPAA," health information held by a covered entity),[31] the Children's Online Privacy Protection Act ("COPPA," personal information collected about minors),[32] and the Gramm-Leach-Bliley Act ("GLBA," nonpublic personal information collected by banks and financial institutions).[33] While any XAI system must comply with these and all other federal laws on data protection, these laws do not contain provisions specific to the AI space.

There has been more activity at the state level, with a number of states adopting comprehensive legislation governing data protection. One of the most prominent state laws is the California Consumer Privacy Act ("CCPA"), which went into effect on January 1, 2020. Similar to the GDPR, the CCPA protects the data privacy rights of California residents. Unlike the GDPR, which applies to the "processing" of personal data, obligations under the CCPA apply to the "collecting" and/or "selling" of personal information by businesses.[34] This does include collecting "by any means" or selling "by electronic or other means,"[35] and therefore can apply to collecting or selling using an AI system. While the CCPA does define the term "processing" as "any operation or set of operations that are performed on personal data or on sets of personal data, whether or not by automated means,"[36] the term comes into play only when identifying which companies must comply with the CCPA.

Specifically, a company, whether or not based or physically present in California, must comply with the CCPA if it does business in California, collects (or has collected on its behalf) personal data, determines (alone or jointly with others) the purposes and means of processing personal information, and meets one of the following: (1) its annual revenue exceeds $25 million; (2) alone or in combination, it annually buys, receives, sells or shares for commercial purposes the personal information of at least 50,000 California consumers, households or devices; or (3) it derives at least half of its annual revenues from selling consumers' personal information.[37] As such, if a company uses an AI system to process personal data, the CCPA may apply (subject to the company meeting the other parts of the definition of a "business" under the CCPA). However, the CCPA does not provide a similar right to "meaningful information about the logic involved" in the case of automated decision-making like the GDPR.

The California Privacy Rights Act ("CPRA"), passed in 2020, amends and expands the CCPA and will become effective on January 1, 2023. In contrast to its predecessor, the CPRA addresses AI, but does not actually create obligations and rights regarding automated decision-making. Instead, the CPRA, using language borrowed from the GDPR, authorizes the attorney general to "[i]ssu[e] regulations governing access and opt-out rights with respect to businesses' use of automated decision-making technology, including profiling and requiring businesses' response to access requests to include meaningful information about the logic involved in those decision-making processes, as well as a description of the likely outcome of the process with respect to the consumer."[38]

The CPRA defines "profiling" as "any form of automated processing of personal information . . . to evaluate certain personal aspects relating to a natural person and in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location, or movements."[39] As such, the obligations eventually imposed by the CPRA could range from being very few to being very similar to those of the GDPR. In fact, the obligations eventually imposed by the CPRA could even be stricter than those of the GDPR. Under Article 22 of the GDPR, automated decision-making, including profiling, is subject to the GDPR's obligations if it "produces legal effects concerning" individuals "or similarly significant affects" individuals. The CPRA does not contain the same "legal effects" qualifier for automated decision-making.

Another prominent state law is the New York Stop Hacks and Improve Electronic Data Security Act ("SHIELD Act"), which imposes data security requirements that went into effect on March 21, 2020.[40] These obligations under the SHIELD Act require that any person or business that "owns or licenses computerized data which includes private information" of a New York resident to "develop, implement and maintain reasonable safeguards to protect the security, confidentiality and integrity of the private information."[41] However, the SHIELD Act does contain obligations specific to the AI space, including any right to "meaningful information about the logic involved" in the case of automated decision-making like the GDPR.

In China, the issue of XAI has also caught the attention of tech watchers and regulators. In March 2020, *Science and Technology Daily*, the official newspaper of the PRC Ministry of Science and Technology, published an article on XAI, the first of its kind as far as we know.[42] The article cites Harry Shum, former executive vice president of Artificial Intelligence and Research at Microsoft, and argues that transparency and explainability are prerequisites for the application of deep learning in AI.

Soon after, the notion of transparency, particularly in automated decision-making that directly affects human users, has found its way into legislation. Modelled after the European GDPR, the Personal Information Protection Law (the "PIPL," effective as of November 1, 2021) has several provisions related to XAI. For example, the PIPL stipulates that, as a general principle, automated decision-making process must "be transparent and generate fair results," and that "unfair discrimination in terms of prices and other transaction conditions" is forbidden. In particular, the law makes clear that "if a decision to be made via automated process has a significant impact on an individual's rights or interests, such individual has the right to request an explanation from the personal information processing entity and prohibit the personal information processing entity from making such decision solely on the basis of automated decision-making."[43] The explanation here would cater to the non-specialist, ordinary user of AI technology.

The law also requires processors of personal information to "conduct prior risk assessments" when, *inter alia*, "using personal information to conduct automated decision-making."[44] Such an assessment would not be possible if the decision-making process takes place in a black box, and therefore an element of explainability is implicit in the clause. However, the target audience would be public and private parties who possess expert knowledge on data science; therefore, the standard here is lower than in Article 24. In addition, a supplier of AI technology shall also "provide [users] with the option not to be targeted on the basis of personal characteristics or provide convenient ways for individuals to refuse."[45] This means that if individual users are not convinced of the use of AI or, if the process is not sufficiently transparent, they have the right to opt out. To what extent such a requirement is a meaningful check on non-consensual AI application remains to be seen, but the inclusion of such language in the PIPL signals the PRC authority's attitude toward AI usage with more transparency and more human agency.

Apart from the PIPL, guidelines from regulatory agencies also confirm increased awareness about XAI. The National Information Security Standardization Technical Committee (formed under the guidance of the national Standardization Administration, which answers to the State Council) issued a list of Normative Guidelines on AI Ethics (the "AI Guidelines") in January 2021. The AI Guidelines set out the protection of "basic rights," including rights to personal safety and privacy, as a "fundamental requirement" of AI activities.[46] In particular, according to the AI Guidelines, researchers and developers should avoid application scenarios where AI technology may be appropriated for illicit and unethical uses—and, presumably to this end, "continuously improve the explainability and controllability of AI."[47] As a practical matter, the AI Guidelines provide that designers and manufacturers shall "describe and *explain* the functions, limitations, security risks and potential implications of an AI system, product or service . . . in a timely, accurate, comprehensive, *clear and unambiguous* manner" (emphasis added).[48] In turn, parties who deploy AI technology should provide end users with the above information along with explanations of "relevant application processes and consequences."[49] In line with the PIPL, the AI Guidelines also explicitly require suppliers of AI technology to allow users to opt out.

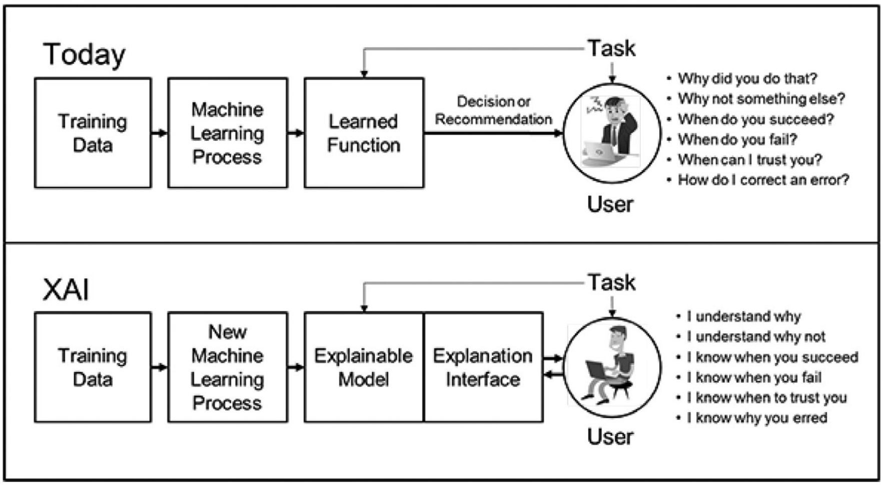
## Explanation of XAI

### What Is XAI and Why Do We Need It?

The U.S. Defense Advanced Research Project Agency ("DARPA") elegantly defines XAI as the ability of machines to explain their rationale, to characterize the strengths and weaknesses of a particular decision-making process and convey a sense of how they will behave in the future.[50] (See Figure 1.)

One of the current issues surrounding both AI in general and XAI is that machines with high levels of learning performance tend to be less explainable. The application of deep learning algorithms to AI-driven models greatly improves learning performance, but this comes at the sacrifice of explainability. As prediction models get

Figure 1. DARPA's XAI Concept. The models will be combined with state-of-the-art human-computer interface techniques capable of translating models into understandable and useful explanation dialogues for the end user.



more accurate their interpretability gets more complex, as accuracy increases, explainability decreases. Thus, one of the aims of XAI is to create machine learning techniques that have increased explainability while simultaneously not compromising on performance.

All technological systems fail at some point, which can involve making relatively simple and embarrassing AI mistakes, as in the case of Microsoft's chatbot Tay, designed to have casual conversations with millennials on Twitter. The chatbot was manipulated by Twitter users into making offensive sexist and racist comments. More serious AI mistakes, however, can result in loss of life. In 2018, an Uber self-driving car struck a pedestrian crossing the road. Analysis of the crash showed that the car detected the pedestrian with her bicycle, but it was not able to correctly identify her as a human being, nor did it accurately predict her path. As advances in machine learning and AI continue, some form of explainability will be necessary not only to comply with (possible) new regulations but also to ensure effective management of the AI. Gartner's 2020 Hype Cycle for Emerging Technologies identified the need for algorithmic trust, stating that:

> Trust models based on responsible authorities are being replaced by algorithmic trust models to ensure privacy and security of data, source of assets and identity of individuals and things. Algorithmic trust helps to ensure that organizations will not be exposed to the risk and costs of losing the trust of their customers, employees and partners. Emerging technologies tied to algorithmic trust include secure access service edge (SASE), differential privacy, authenticated provenance, bring your own identity, responsible AI and XAI.

## What Does It Mean to Explain AI?

In an article published in 2017, Lipton described several properties of interpretable models which relate to two broad questions: how does the model work and what else can the model tell me?[51] Establishing how the model works could be viewed as its transparency, which Lipton breaks down into three levels: the entire model (simulatability), the individual components (decomposability), and the training algorithm (algorithmic transparency). According to Lipton, a model would be deemed to be simulatable if the whole is able to be understood by a human user at once. This means that the model's complexity is limited by human understanding. A decomposable model is one in which each component of the model, each input, parameter and calculation is transparent and admits an intuitive explanation. Lipton explains that this could be as simple as each node in a decision tree having a plain text explanation, or the parameters of a linear model being described as representing strengths of association between each feature of the label. Lipton goes on to note that while these strengths of association may seem intuitive they can be greatly affected by both feature selection and pre-processing.

The example given by Lipton is that associations between flu risk and vaccination may appear positive or negative depending on whether the feature set used contained indicators of age or immunodeficiency. Algorithmic transparency refers to how well understood the training process used to develop the model is. It is known that training a linear model will converge to a unique solution. However, the techniques used to train neural networks are not understood, meaning that deep models do not display such algorithmic transparency. It is worth noting that Lipton points out that we, as humans, do not display any of these forms of transparency.

When it comes to explaining AI, an important question that needs to be taken into consideration is "who is asking for an explanation?" Much of the current research into XAI has been focused on the researcher and the developer of the AI system, particularly in the context of a single neural network. However, in order for AI to be trusted for use in wider society, other types of explainability will be required. With regard to the AI present in AV, there are a number of points of interaction between people and autonomy, thus the explanation methods should be appropriate for the type of interaction.[52] In a 2019 paper, Glomsrud et al. suggest that there are four types of explanation required for AV. These explanations should be tailored toward the developer, assurance, the end-user and external explanation requirements. When considering the explainability of a system, it is important to consider what is the purpose of the explanation and who is it tailored toward, that is, how complex and in-depth the explanation needs to be. For example, a researcher/developer of an AI model will be able to understand and will require a more detailed explanation than the owner/user of an AV.

It is also important to consider whether the explanation needs to be given in real time or retrospectively. For developers and those considering cases of liability, the explanation of an AI system's decisions can be retrospective. However, a passenger in the AV or a pedestrian may require some form of an explanation in real time. An example of a real time explanation that assists pedestrians is the "smiling car concept" designed by Semcon AB.[53] The car "smiles" to show pedestrians that it has noticed them.

## How Do We Explain AI?

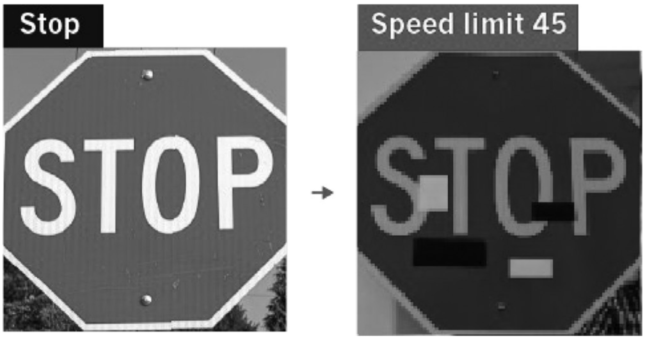
We can think of AI as being an umbrella term under which the subsets of machine learning and deep learning sit. AI is a technique that enables machines to mimic human behavior. If we then think of machine learning as being a subset of AI that utilizes statistical methods to enable machines to improve over time with experience, then deep learning is a further subset of machine learning that is based on how the neurons of the human brain function, leading to the term "artificial neural network." These neural networks are the "black boxes" that XAI is trying to explain.

We can describe machine learning in general, as follows.

Training data is used to train a model in a learning algorithm. Once complete, this results in a learned function ("model"). This model can then be fed inputs that results in an output/prediction function. This output is what the user sees. The machine itself decides which aspects of the input to pay attention to and which to ignore. These machine learning systems are capable of analyzing large amounts of data to create output conclusions/predictions; however, they do not give any justification for the output.[54] If we think of the input as an image and the output as a classification, XAI is trying to explain why the image was classified in that particular way and which variables played an important role in how the machine classified the image.

The following example highlights why it is important to understand why an image is classified a certain way, rather than simply assume that because the image is classified correctly that the machine is identifying and learning from the correct features. Take a model that learned to distinguish between dogs and wolves. However, the model did not actually learn the differences between dogs and wolves but instead learned that the wolves had snow in their picture while the dogs had grass. This is an example of a dataset that contained implicit bias within it and the algorithm

Figure 2. The addition of stickers to this road sign confused an AI System into reading a "stop" sign as a "speed limit 45" sign.



tuned into that. This kind of bias and others similar to it may not be noticeable to humans, and is therefore very hard to correct without understanding how a model makes a decision.

Neural networks have a set of input units through which raw data is fed; the inputs are mapped to output units, which then classify the input data. Deep learning neural networks contain several hidden layers in between the input and output units. The raw data is fed into the input units, which then trigger some of the units in the next layer, and then the next according to mathematical rules until they reach the output units, which then give a classification. The multiple, hidden layers result in deep learning neural networks that are theoretically capable of complex classifications of data. These neural networks have proven to be able to classify images that fall within the realms of their training data with high accuracy; however, they can be easily fooled by simple, small changes to the input data, such as sections of white noise inserted into audio inputs or sticking rectangular blocks onto road signs (Figure 2). Understanding why these neural networks are so easily fooled in certain situations is key to improving the reliability of AVs.

According to Gilpin et al. 2018,[55] there are currently two broad approaches being taken to make AI explainable: designing the AI model to be inherently explainable, and using an external model to explain a first AI model. The first approach requires that the AI system itself is able to be understood by a human user; this may have the drawback of the system being of a reduced complexity. The second method will rely on post hoc explainability, and will involve extracting some form of explanation from the highly complex models.[56] Post hoc explanations are categorized as either backpropagation-based methods or perturbation-based methods. Backpropagation models work by calculating the gradient (change in weight with regard to change in error) of a model's target output to the input; thus, they identify the contribution of each input feature to the output using a few backward passes through the network. Perturbation methods work by perturbing or removing specific input features and measuring what outcome this has on the output.[57] Models can also be split into local explanations and global explanations. Local explanations aim to explain a model's single prediction, while global explanations aim to explain a model's decision-making process in general.

There are a number of methods of addressing different aspects of machine learning in order to work toward the overall goal of XAI. If we consider in the case of an AV that image classification and computer vision play important roles, the AV has to be able to identify a huge variety of different objects, from interpreting road signs and traffic signals to differentiating between a pedestrian, a cyclist and a vehicle. Some examples of XAI involved in computer vision include:

- Local Interpretable Model-Agnostic Explanations ("LIME");
- Gradient-weighted Class Activation Mapping ("Grad-CAM");
- Randomized Input Sampling for Explanations ("RISE"); and
- Textual Explanation for Self-Driving Cars.

### *LIME*

LIME as the name suggests, is used to explain individual predictions, it is a post hoc model agnostic method of interpretability and aims to explain the predictions of any black box learning model. As LIME is model agnostic, it is not able to explain what goes on inside the model itself, but rather focuses on perturbing the input and measuring how the model's predictions change. The perturbed data points are then weighted based on how close they are to the original example.

### *Grad-CAM*

Grad-CAM is a post hoc attention for producing heat maps that is applied to an already trained convolutional neural network, [plus] Grad-CAM is able to be used on any model. The aim of Grad-CAM is to understand at which parts of an image a convolutional layer "looks" for a certain classification. We feed an image into the network to calculate the Grad-CAM heat map for that image for a chosen class of interest, it is only interested in the features that have a positive influence on the class of interest. The technique is complex however the output is intuitive.

### *RISE*

RISE was developed at Boston University and aims to explain image classification decisions by generating a saliency importance map for each decision. RISE uses masked versions of images to identify which pixels in each image correspond to the maximum importance for each prediction. It does not use internal information from the neural network (e.g., output layer gradients or the weighting of neurons). Like LIME, RISE can work with any type of pre-trained black box model.

### *Textual Explanation for Self-Driving Cars*

Work done by the University of California, Berkeley, used visual attention to identify the behavior taken by a self-driving car and then suggested an explanation for the behavior. The aim of this work was to provide a "natural language textual explanation," for example, "vehicle slows down" and "because it is approaching an intersection and the light is red." One of the major advantages of a natural language explanation is that it is inherently understandable to humans and does not require and understanding of the design of the AI system. This model was trained using explanations from human annotators, and building an explanation dataset that can then be built on top of another driving dataset. The driving dataset uses video captured from dashboard cameras in human driven vehicles. The model proposed in the paper explains how a driving decision was made by visualizing the areas of the image that the decision maker is focusing on and by generating a textual description and explanation of what has caused a particular decision.

## Liability

As will be discussed below, the information asymmetry that may exist between an AI developer and user may be bridged by

XAI. However, the complexity of certain AI systems is such that the AI developer is not able to trace the exact cause. In such circumstances, XAI may also offer a solution. A way to circumvent issues of attributing cause is to implement a strict liability regime. Whether any strict liability regime will indeed be implemented in relation to AI is still uncertain.

It has also been argued that current liability regimes are in some cases adequate to cover advanced AI systems,[58] which follows the reasoning that open legal terms such as duty of care are flexible enough to be applied to AI. If a producer of AI is held responsible for its AI, the producer must show it has taken adequate safeguards to fulfil its duty of care. In the automotive sector, cars have been controlled by software for decades (not called "AI"). For years cars have been equipped with sensors that provide one or more processors with data on engine conditions and traction/grip, which the processor analyses and uses to modify what the engine and transmission are doing, for example, to alter fuel injection rates and boost pressure, and to run the traction control and anti-lock braking systems. A defect in any one of those software systems can easily result in significant physical injury or death.[59] Liability for failure of software has been generally established.[60] Given that AI is also software, why/how should it be treated differently? In 2020, the European Commission neatly summed up the issue of establishing liability in the case of AI:

> Due to the black-box effect in some AI, getting compensation could become difficult for damage caused by autonomous AI-applications. The need to understand the algorithm [code] and the data [dataset] used by the AI requires analytical capacity and technical expertise that victims could find prohibitively costly. In addition, access to the algorithm and the data could be impossible without the cooperation of the potentially liable party. In practice, victims may thus not be able to make a liability claim. In addition, it would be unclear, how to demonstrate the fault of an AI acting autonomously [bad processing → producing], or what would be considered the fault of a person relying on the use of AI [bad input → operator].[61]

In 2017, the European Parliament noted that the complexity and the capacity for self-learning and the potential autonomy of AI systems, as well as the multitude of actors involved in the AI supply chain, represent challenges to the European liability regimes.[62] The Parliament argued that corrections to the liability regimes are necessary to avoid a situation in which claimants end up without reparation. The European Parliament's 2017 report on "Civil Law Rules on Robotics" called for a liability regime for certain new technologies such as robots.[63] The European Commission's Expert Group on Liability and New Technologies ("New Technologies Formation Report") proposed strict liability on both the producer and operator of an AI application. Strict liability means that operators or producers of high-risk AI systems could be held liable for any damages caused by an autonomous activity, device, or process driven by their AI system, even if the operator or producer did not act negligently.

The New Technologies Formation Report notes that any high-risk AI operator, for example, AI-driven robots in public spaces, should be subject to strict liability for damages resulting from its operation. The New Technologies Formation Report illustrates that if the sensors controlling the path of an AV malfunction, causing the AV operated by A to leave its intended path and run into a pedestrian B on the street, B should be able to seek compensation from A without having to prove that A is at fault. Introducing the risks of a high-risk AI application thus carries the risk of being held liable for any damages resulting from such high-risk AI application.

The Report also proposed a fault-based liability system for non-high-risk AI applications: depending on the fault of the operator or producer, they can be held liable. Any person using technology that does not pose an increased risk of harm to others would then still be required to ensure proper safeguards for the technology and should be liable for breach of such duties if at fault. One could question if it is fair that an operator can be held liable if the AV changed lanes without input from the operator. This illustrates the need to distinguish between the actions of the AI's end user [bad input] and the actions of the AI [bad processing]. Assuming that the instructions for use, and warnings about misuse, are sufficient, the producer should be held liable only for bad processing.

The New Technologies Formation Report further states that producers of technologies should be required to log information concerning the operation of the technology: "logging by design."[64] Logging data is meant to reconstruct events and causal chains to allocate liability (for example, by finding out which AV has caused the crash by not replying to a signal sent by the other AV).[65] The

Report states that "the absence of logged information or failure to give the victim reasonable access to the information should trigger a rebuttable presumption that the condition of liability to be proven by the missing information is fulfilled." The Report specifically points to AVs as a case where a logging duty exists:[66]

> Take the example of a crash between A's AV and B's AV, injuring B. The traffic situation was one where, normally, the two AVs would exchange data and "negotiate" which AV enters the lane first. When sued by B, A refuses to disclose the data logged in her AV's recordings. It is therefore presumed that her AV sent a signal telling B's AV to enter the lane first, but nevertheless went first itself.[67]

The logging requirement has become part of the AI Regulation proposal presented by the European Commission in April 2021. With regard to the burden of proving causation, the Report notes that in general it is the victim that should prove fault, but such burden may be alleviated in case of informational asymmetry.[68] To support the analysis of the cause of damages, the victim should be provided with information about the root cause, including the degree of processes within the technology that may have contributed to the cause. The above does provide a clear example of where XAI will be beneficial. However, Bertolini argued that the logging by design proposal is problematic because of the "interpretation and analysis that might be extremely complicated and costly. Moreover, even the most accurate ascertainment of the dynamics of the event might be insufficient to establish liability."[69]

The European Commission published a white paper on Artificial Intelligence on February 19, 2020, together with an accompanying Report on the safety and liability framework (the "White Paper Accompanying Report"). In the White Paper Accompanying Report, the Commission points to the difficulty of tracing back potentially problematic decisions taken by AI systems. It holds that under the regular product liability regime,[70] a manufacturer is liable for damage caused by a faulty product, but in the cases of autonomous vehicles: "it may be difficult to prove that there is a defect in the product, the damage that has occurred and the causal link between the two."

In the white paper, the Commission is interestingly silent on the option of implementing a strict liability regime. The white paper proposes a risk-based approach: differentiating between high-risk AI and low-risk AI and appropriate regulatory intervention for each. It does refer to the lack of transparency of AI tools, and refers to the Report, but critically, the white paper does not refer explicitly to implementing strict liability. In the final report that summarizes the responses from the public consultation on the basis of the white paper,[71] it is noted that, "concerning liability, many business associations and large companies thought that existing rules were probably already sufficient or they were sceptical of strict liability rules and possible regulatory burdens."[72]

In October 2020, the European Parliament adopted a proposal on liability (the "Liability Proposal"), in which it calls on the European Commission to present a proposal for a regulation based on the EP's proposals. The Liability Proposal echoes the earlier reports and suggests implementing strict liability for operators of high-risk AI and defining a list of high-risk AI applications, which includes all AI operating in public spaces, including AVs. The Liability Proposal notes that all AI systems may technically be the direct or indirect cause of harm or damage, yet are nearly always the result of someone's programming systems. The black box element of AI could make it difficult to attribute cause to human input or to decisions in the programming. The European Parliament notes that, "by making the different persons in the whole value chain who create, maintain or control the risk associated with the AI-system" liable, the black box issue is circumvented.

As Zech argues, it is a misconception to only view strict liability from an incentive perspective as detrimental to innovation.[73] Rather, he argues, strict liability provides an incentive to further develop technologies to make them safer. As such, strict liability's benefit as a policy option is that the producer is incentivized to develop safe and accurate AI. Therefore, strict liability may further install trust. If the purpose of implementing XAI is to install public trust in the technology, and it is true that there is a trade-off between explainability and accuracy, it is reasonable to question whether implementing strict liability is more effective than mere transparency requirements. However, as Zech argues, transparency may complement liability in the sense of provability.[74]

Like the European Union, current tort liability regimes in the United States—strict liability and negligence—could similarly be applied to an AI system, with the same potential arguments for and against as above. Unlike the European Union, the U.S. government has not issued as many detailed proposals or reports on potential liability schemes specific to AI. In early 2020, the Office of Management and Budget issued a memorandum on "Guidance for Regulation of Artificial Intelligence Applications" (the "OMB Guidance"). The OMB Guidance "sets out policy considerations that should guide ... regulatory and non-regulatory oversight of AI applications developed and deployed outside of the Federal government."[75]

One such policy consideration is a cost-benefit analysis when agencies consider the application and deployment of AI into already regulated industries, recognizing that the introduction of AI may "create unique challenges. For example, while the broader legal environment already applies to AI applications, the application of existing law to questions of responsibility and liability for decisions made by AI could be unclear in some instances."[76] As such, agencies need to, consistent with their authorities, "evaluate the benefits, costs, and distributional effects associated with any identified or expected method for accountability."[77]

Similarly, in China, there is a rising interest in the issue of tort liability in the context of AI, especially autonomous vehicles. Being a hub of technological innovation, the southern city of Shenzhen plays a pioneering role in regulating AV in China. The municipal government's *Proposed Regulations on Smart Net-Connected Cars* ("Shenzhen Proposal") published in early 2021 devoted a chapter to the identification and allocation of liabilities in cases of AV traffic accidents (chapter 8). The regime is broadly consistent with practices in other parts of the world where the operator, or human driver, is directly liable for accidents in which human input is present. Such a result is straightforward if the accident arose proximately from human input errors, as the driver is held exclusively responsible just like she would be in an accident that did not involve AI. However, if the actual cause of the accident is defects (either in design or in manufacturing) inherent in the AV, the driver would rightly consider it unfair to have to bear responsibility for things over which she has no control or even understanding of. Therefore, the Shenzhen Proposal provides for a channel whereby the human operator has a recourse against the producers and sellers of smart cars.[78]

In cases where the AV is advanced enough that no human input is required, the Shenzhen Proposal identifies the directly liable parties as the controller or owner of such cars. The implication is that, in acquiring the AV hardware and software, the controller or owner assumes all associated risks, including the risk of programming or design defects. Similar to the case with human drivers, controllers and owners may also claim against the producers and sellers.[79] The Shenzhen Proposal further allocates liabilities between controllers and owners. In particular, a controller would be liable for traffic violations and accidents that occur when an autonomous driving system is in full control of or taking over the dynamic driving of a vehicle, and an owner would be liable for the traffic violations and accidents in all other circumstances.[80] On-vehicle devices, roadside devices, and video surveillance recordings are all acceptable as circumstantial evidence for liability in accidents. [81] Notably, the Shenzhen Proposal makes human operators directly liable, but unlike either, it also adds sellers of smart cars to liable parties, along with manufacturers. In general, the Shenzhen Proposal provides a moderate and pragmatic solution to the problem that is analogous to and consistent with existing doctrines of product liability and shuns radical notions such as legal personhood for robots.

It is worth noting that the issue of AI liability has stirred broader discussions in China. For instance, a widely cited article by Liu[82] enumerates the drawbacks of decision-making by smart car.[83] According to Liu, even if AI technology has been honed to overcome such restraints, the problem of explainability persists. As Liu observes, the lack of transparency has a bearing on the determination of AV liability in road accidents, and models that are based more on probabilities rather than causal logic may also upend conventional knowledge about what exactly constitutes an act of fault or negligence. According to Liu, who denies AI of moral agency, liability naturally falls upon human actors such as the producer or operator. Given that AI is at best glorified tools under this view, ordinary product liability rules would apply with only a few modifications.[84] Liu believes that, if the human driver is aware of the risk of AI but uses it anyway, she should be held liable for negligence in the event of an accident.

In contrast, Yang[85] argues that the liability of any AV—no matter "fully automated" or "highly automated with human input"—should be traced to the "vehicle" but not its owner or operator; in practice this means that "the vehicle's designer and manufacturer" should be held liable by default, unless there is evidence that the error took place under the full control of the operator. Yang does make an exception to producer liability in cases where "defects were not detectable by the level of technology available at the time of the product's circulation"—the basis for this exception is Article 41, Clause 2 of the PRC's Product Liability Law. Yang also considers cases where accidents arise as a result of malicious cyberattacks; in such cases the culprits behind the attacks, aka hackers, should be properly found liable.

Standing on another extreme, Zhang and Xiao[86] present the argument that technological advancement and rights theory naturally lead to AI's legal personhood independent of its makers. As such, the framework over AI liabilities must differ from existing ones over product liability. Zhang and Xiao find Yang's solution naïve and detrimental to innovation, since technological research and development more often than not involve unknown risks. However, they agree with Yang that the human operators should not be held primarily liable, if at all; but interestingly, their rationale for exculpating drivers is not bad programming, but rather the "learning and adapting ability of the AI" itself. Citing cases where robots are being granted citizenship and resident rights in Saudi Arabia and Japan, Zhang and Xiao argue that AI liability may not be as outlandish as it first appears; however, they are silent on how exactly this may be realized with current technology. Instead, Zhang and Xiao propose a regime of collective enterprise liability for industry players in cases where AVs are used by the public. Perhaps as an intermediary measure, they also acknowledge the liabilities of producers and manufacturers, but with different standards: design defects should be held to a negligence standard in order not to stifle innovation, whereas production defects should be held to a strict liability standard.

## AI Regulatory Requirements

The regulatory trend of governing AI in the European Union is clearly pointing toward transparency requirements. On April 21, 2021, the European Commission published its Proposal for a Regulation on a European approach for Artificial Intelligence (the "AI Regulation"). The AI Regulation contains far-reaching transparency rules. Recital 47 states that "to address the opacity that may make certain AI systems incomprehensible to, or too complex for natural persons, a certain degree of transparency should be required for high-risk AI systems. Users should be able to interpret the system output and use it appropriately."

The AI Regulation contains far-reaching requirements for high-risk AI systems. For example, the AI Regulation sets out specific requirements for datasets (the input of an AI system) to prevent bias and discrimination. Datasets must be "error-free, representative and complete." However, it is unclear what is meant by "error-free" datasets, especially since datasets usually are not error-free, regardless of which definition is used.[87] Recital 44 AI Regulation qualifies that datasets must be "sufficiently" error-free—a 5.5 out of 10 is suffices? Another example is the requirement for human supervision. The AI Regulation states that the provider must allow human supervision of the AI system to minimise risks to health, safety or fundamental rights, by a person who fully understands the capabilities and limitations of the system and can decide not to use the system or its output. The AI Regulation places an important requirement on providers to ensure transparency, so that the user of the system can interpret the output of the AI system and use it appropriately. The draft AI Regulation contains a requirement to register high risk AI system technical details in a public register. Furthermore, Article 14 contains a requirement for human oversight, which should enable a human to fully understand the capacities and limitations of the high-risk AI system. Article 62 contains an obligation notify the authorities if a causal link is identified between the AI system and a serious incident or any malfunctioning of the AI. In its current version, the draft AI Regulation will therefore require some form of XAI.

Transparency obligations are not the only requirement the AI Regulation contains. The AI Regulation also requires accurate AI systems. The sticking point here is the trade-off between accuracy and explainability, which has been raised by several scholars.[88] The more accurate the AI, the less understandable it is for a human. Transparency of AI is a hot topic. The cliché that AI is a black box application where something "magical" happens to data and then output comes out, is not necessarily true for every application of AI. Some AI applications are simple and easier to understand (if this, then that), but that again depends on the definition of AI. More complex forms of AI are inscrutable and less easy to explain.

However, the AI Regulation requires high-risk AI systems to be capable of interpreting the output of the AI system. The point of the AI Regulation is thus that it requires providers to provide both an "appropriate level of accuracy" and human-interpretable AI systems. And all this with error-free datasets. On the one hand, it is not unreasonable to set high requirements for "high risk AI" and transparency is an important requirement anyway according to the Court of The Hague;[89] on the other hand, it is unclear how providers should comply with these obligations.

The AI Regulation contains further specific transparency requirements for other sorts of AI applications. This includes AI systems that can interact with individuals (e.g., chatbots), detect emotions or determine association with social categories based on biometrics, or generate or manipulate content that appreciatively resembles authentic content (e.g., deepfakes).

Noncompliance with the AI Regulation may generally be fined up to €30,000,000 or six percent of the annual worldwide turnover, whichever is higher.

Prior to the publication of the AI Regulation, on December 15, 2020, the European Commission published a proposal for a regulation on a single market for digital services—the Digital Services Act (the "DSA"), which also contains AI related transparency obligations. For each advertisement, the online platforms must provide, in real time, clear and unambiguous information to each user about the main parameters used to determine why a specific user is targeted by this ad. In addition, the DSA refers to "recommender systems" defined as a "fully or partially automated system used by an online platform to suggest in its online interface specific information to recipients of the service, including as a result of a search initiated by the recipient or otherwise determining the relative order or prominence of information displayed."

Under Article 29 (1) DSA, the recipients of the service have the right to know the main parameters of recommender systems, as well as having options to influence/modify those parameters. While a legal requirement to transparency is proposed in both the draft DSA and the draft AI Regulation, the draft AI Regulation goes as far as requiring a human to fully understand the capacities and limitations of the high-risk AI system. As such, the draft AI Regulation's regulation may require AI to be explainable in the sense of offering an explanation for high-risk AI system processes. Any fault-based liability regime will require XAI to ensure cause can be attributed. Regulatory AI proposals currently focus on logging details of the AI. These explanations safeguard that AI systems can be reviewed for whether the decisions they made are accurate. Such legal requirements also help developers to better comprehend the AI system.[90]

Like the European Union, the federal regulatory trend of governing AI in the U.S. is pointing toward transparency requirements. In 2019, both chambers of the U.S. federal government proposed the Algorithmic Accountability Act (the "AAA"). While the AAA ultimately was not passed, it would have empowered the Federal Trade Commission ("FTC") to require entities that use, store, or share personal information to conduct "automated decision system impact assessments" and data protection impact assessments. The AAA would have applied to any entity that is subject to the FTC's jurisdiction and makes more than $50 million per year; possesses or controls personal information on more than one million consumers or consumer devices, or primarily acts as a data broker that buys and sells consumer data. The required impact assessments would have involved evaluating AI systems used by the entity "for impacts on accuracy, fairness, bias, discrimination, privacy, and security."[91]

Unlike the AAA, the National AI Initiative Act of 2020 (the "NAII Act"), passed on January 1, 2021, does not contemplate specific AI technology regulations, but creates a framework for coordinating government agencies in the creation and implementation of the United States' overall AI strategy, including potential future regulation of the technology. The NAII Act establishes a federal initiative to accelerate and coordinate federal investments and facilitate new public-private partnerships in research, standards, and education in AI.[92]

Other U.S. federal agencies have published guidelines. In April 2020, the FTC published "Using Artificial Intelligence and Algorithms" that outlines best practices for businesses in the United States utilizing AI technology, including: (1) being transparent by allowing consumers to know when AI technology was utilized and when such technology was collecting sensitive data; (2) informing consumers as to the reasons why a good or service was denied using AI technology; and (3) ensuring implemented AI technology makes fair, non-discriminating decisions.[93] Additionally, in 2021, the FTC published "Aiming for Truth, Fairness, and Equity in Your Company's Use of AI." Although this article does not discuss the creation of new legislation, it highlights that the FTC is equipped and prepared to address issues that may arise from the use of AI technologies in commerce.

Specifically, the article emphasizes: (1) Section 5 of the FTC Act, which prohibits unfair or deceptive practices, including the sale or use of, for example, racially biased algorithms; (2) the Fair Credit Reporting Act, which comes into play in certain circumstances where an algorithm is used to deny people employment, housing, credit, insurance, or other benefits; and (3) the Equal Credit Opportunity Act, which makes it illegal for a company to use a biased algorithm that results in credit discrimination on the basis of race, color, religion, national origin, sex, marital status, age, or because a person receives public assistance.[94]

The National Security Commission on Artificial Intelligence (the "NSCAI") was created in 2019 with the task of making recommendations to the President and Congress to "advance the development of artificial intelligence, machine learning, and associated technologies to comprehensively address the national security and defense needs of the United States."[95] The NSCAI released its final report in March of 2021 (the "NSCAI Final Report"), which "presents an integrated national strategy to reorganize the government, reorient the nation, and rally our closest allies and partners to defend and compete in the coming era of AI-accelerated competition and conflict" through 16 chapters that provide topline recommendations.[96] Regarding federal regulation, chapter 8 of the NSCAI Final Report addresses the impact of AI technology on privacy, civil liberties, and civil rights in the United States. At the conclusion of the chapter, the NSCAI proposes the following steps for the U.S. federal government:

1. Invest and adopt AI tools to enhance oversight and auditing in support of privacy and civil liberties;
2. Improve public transparency about how the government uses AI;
3. Develop and test systems with the goal of advancing privacy preservation and fairness;
4. Strengthen the ability of those impacted by government actions involving AI to seek redress and have due process; and
5. Strengthen oversight mechanisms to address current and evolving concerns.[97]

In China, there is currently no existing or proposed law specific to AI. That said, the AI Guidelines mentioned above, though nonbinding, may help to provide a preliminary sense of the regulatory bodies' approach as to the governing the ethics of development and application of AI technology. We anticipate more systematic regulation in the area will be supplemented by future legislation.

## Legal Implications of XAI

As XAI may be required for certain AI applications, it is important to consider the legal implications of XAI. Two important elements should be distinguished: (1) legitimate reasons to limit transparency, and (2) a more fundamental issue, the development of a standard of explainability.

### Legitimate Reasons to Limit Transparency

If the purpose of XAI is to provide transparency and explain its logic, how do we protect this? If XAI is to be truly used, then logic dictates that the AI used in each aspect of the vehicle is explainable. If some aspects of this are protected by trade secrets—for example, the visual attention technology responding to outside stimuli—would this be available to the final manufacturer of the vehicle? An XAI machine is unlikely to be able to distinguish between types of information; using XAI could then potentially lead to the disclosure of commercially sensitive information.[98]

Most protection with regard to AI comes from trade secrets.[99] According to the World Trade Organisation Agreement on Trade-Related Aspects of Intellectual Property Rights ("TRIPS"), and is generally consistent with the various national laws around the world, trade secrets allow the prevention of disclosure of information that is secret, has commercial value and has been subject to steps to keep it secret.[100] As Meyer argues, businesses have a legitimate interest to protect their algorithms from being known publicly from a commercial perspective, which counterbalances the requirement for transparency.[101] However, the very nature of XAI algorithms may make maintaining such information as a trade secret more difficult. Indeed, the EU Trade Secrets Directive[102] provides for an exception for the purpose of protecting a legitimate interest recognized by EU or national law. According to Brkan, explaining AI to a data subject could fall under this exception as it is provided by the GDPR and seeks to protect a legitimate interest. This weighing in favor of secrecy may be particularly difficult to assert when it comes to health and safety and regulation in relation to AV. The right to an explanation, continues to be a controversial topic.

As Zech argued, IP protection may act as a counterbalance for transparency obligations.[103] In this respect, the draft AI Regulation considers intellectual property rights and limiting disclosure to authorities. According to paragraph 3.5 of the draft AI Regulation's impact assessment, disclosure will be limited. Any disclosure of information will be carried out in compliance with relevant legislation in the field. When regulators need to be given access to confidential information or source code to examine compliance with substantial obligations, they are placed under binding confidentiality obligations. In order to provide for such transparency within the GDPR, the source code, or even the way the algorithm operates, does not necessarily need to be disclosed.

Another legitimate concern in the case of XAI is privacy. As the New Technologies Formation Report stated, logging might be in some cases inappropriate. The report illustrated this by means of a dystopic scenario of an AI-equipped doll for children, where the risks associated with the doll do not outweigh the negative implications of logging for data protection reasons, because logging in this context means recording the children).[104] However, in the case of an AV, the report does indicate that the risks of the AV are such that logging would be suitable. Depending on the type of AI, privacy may be a counterbalance to transparency.

Another implication of XAI is that it may expose the risk factors used in the model. If an XAI system is able to disclose which aspects of an image are most important in determining the output prediction and the algorithm is protected, does this expose proprietary data that would render the AI ineffective? Depending on what you make AI-assisted decisions about, you may need to protect against the risk that people may game or exploit the AI model if they know too much about the reasons underlying its decisions. For instance, in the case of using AI-assisted decisions to identify wrongdoing or misconduct (e.g., fraud detection), the need to limit the information you provide to individuals will be stronger, particularly in relation to the rationale explanation. However, you should still provide as much information on reasoning and logic as you can. There is a need to avoid abuse, signifying that too much transparency in the wrong context may actually defeat the purpose of an AI system.[105]

There can be incentives for "gaming the system"—examined by de Laat[106] in terms of "perverse effects of disclosure"—affecting everything from trending topics on Twitter to security issues and welfare distribution. If an AI system is more transparent, does this leave it more exposed to manipulation or hacking? For example, if it is known that an AI system in an AV struggles to distinguish between a stop sign and a speed limit sign if stickers are placed on the speed limit sign, this could leave the AV open to malicious manipulation. This is likely to also apply to non-malicious issues too; for example, where signs have been damaged due to weather conditions. However, in other settings, there will be relatively few risks associated with giving people more detail on the reasons for decisions. In fact, it will often help individuals to legitimately adjust their behavior or the choices they make in order to achieve a desirable decision outcome for both parties.

The above shows that there are legitimate interests that in some cases justifiably require a limitation to the transparency of AI. Striking a balance between these interests is a challenge. Whether the current legal requirements to XAI appropriately find a balance needs to be further researched.

## Developing a Standard of Explainability

In the context of AVs, XAI is in its infancy. Most XAI research is focused on a single neural network, rather than a multitude of systems all interacting to allow an AV to make a decision. Explaining an image classification model that differentiates between pedestrians and vehicles is one thing; however, the complexity involved in explaining why an AV swerved into another lane or slowed down is a far greater task.

But what constitutes explaining the AI? We might look to the risk factors in the model: do we understand what each represents? Alternatively, we could consider the model's complexity: is it simple enough to be examined by a human being?[107] Post hoc interpretations may explain an AI machine's prediction but not automatically explain the way it works. "Interpretability" refers to which factors lead to an output, and "explainability" refers to how the mechanism actually works. With regard to the interpretation of how the AI system is using different pieces of data and how that data impacts the machine's decisions, a 2018 study by Poursabzi-Sangdeh et al.[108] found that too much transparency hampers users' ability to detect when the machine makes a mistake and to correct for it. Researchers hypothesise that this is due to information overload.

It is not enough to have all the data on why an AI made a specific decision, that is, the complete decision tree, which may be vast. To make practical use of it, data subjects must be able to identify the key decisions and understand why one was chosen over the other.[109]

At the most basic level, XAI might be implemented by keeping a log of every outcome of every conditional (if this, then that) step in the program, but this is likely to result in torrents of data.[110] The argument is that dumping data is not meaningful for any individual. Taking pictures of every spare part in the car will not tell you why the driver was speeding. Edwards and Veale posed the question: If meaningful information about the logic of AI is so hard to provide, how sure are we that explanations are actually an effective remedy, and if so, to achieve what?[111] However, the point of XAI is to give meaningful information that can be interpreted by humans. As has been argued, it is not necessarily about individual's rights but about the possibility of civil society (e.g., non-governmental organizations and academics) to review the accuracy and potential biases of AI systems.

Nevertheless, as XAI becomes a legal obligation it is important to formulate a clear definition of what it is we are trying to explain and on what level. In principle, it is possible for an AI to explain how it works, and explain how it explains how it works, and explain how it explains how it explains—ad infinitum. Can human decision-making be explained using the logical hard and fast rules that govern mathematics and vice versa? Do we need true transparency, that is, do we need to understand exactly how the neural model works? Or is it enough that the model is interpretable rather than truly explainable? The academic debate on this is still ongoing.[112] Legislators, regulators and academia can step in to standardize XAI and create best practices.

The above does show that even an organization that has implemented XAI is not without troubles. XAI as such creates a new layer of complexity: It must be balanced with other interests, comes with its own risks, and there is no one-size-fits-all solution, let alone a clear definition of explainability.

## Conclusion

This article has reviewed current legal requirements for explaining AI and its legal implications. If an organization is committed to being accountable and transparent about its AI, it is important to assess the risks that come with XAI. Weighing up the scales, there are both benefits and risks associated with deploying XAI. XAI is not a panacea, and certain risks associated with explaining AI identified in this article are inherent to XAI. As discussed, intellectual property considerations may be trumped by legal requirements, and in most cases it does not require disclosing commercial secrets. There are potential risks with disclosing details about data subjects or with gamification. In certain settings this issue may be greater than in others.

There is a need to strike a delicate balance between public interest in transparency and commercial, privacy, and other interests. For example, as noted in the first draft of the guidelines from the EU high-level expert group ("HLEG"), there might be "fundamental tensions between different objectives (transparency can open the door to misuse; identifying and correcting bias might contrast with privacy protections)."[113] There is a good chance that companies producing AIs will have to be able to explain how they work—sooner rather than later in sectors with more regulation. The degree of explanatory detail required may depend on the sector and will be affected by considerations around IP, privacy and security (re: gamification). Whatever sector or business you are in, explaining your AI-assisted decisions to those affected will help to give you (and your board) better assurance of legal compliance, mitigating the risks associated with noncompliance. However, XAI is another layer of complexity that might go wrong. This is just a piece of the solution to build trust, but overreliance could have its own problems.

## Notes

[*] David van Boven (david.vanboven@allenovery.com) is an associate at Allen & Overy in Amsterdam and part of the firm's Tech Team, specializing in data protection, privacy law, and intellectual property. Paul B. Keller (paul.keller@allenovery.com), a member of the Board of Editors of *The Journal of Robotics, Artificial Intelligence & Law*, is a partner in the firm's litigation practice in the New York office, specializing in intellectual property. Harriet Ravenscroft is a member of the firm's internal scientific team. Jill Ge (jill.ge@allenovery.com) is counsel in the firm's office in Shanghai, covering the full spectrum of IP litigation and transactions in China. Wentao Zhai, a third-year student at Harvard Law School, previously was an intern in the firm's Shanghai office. Arwen Zhang, a senior legal counsel at Medtronic, previously

was a China legal specialist at Shanghai Lang Yue Law Firm (Allen & Overy's joint operation partner in China).

[1] House of Lords Liaison Committee "AI in the UK: No Room for Complacency," 7th report of 2019-21, December 18, 2020.

[2] Kay, M., Matuszek, C., and Munson, S.A. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations (Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems ACM, 2015), 3819.

[3] Zuiderveen Borgesius, F.J. (2020). Strengthening legal protection against discrimination by algorithms and artificial intelligence. The International Journal of Human Rights, 24(10), 1572-1593.

[4] Nyholm, S., and Smids, J. The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? Ethical Theory and Moral Practice 19, 1275-1289 (2016). https://link.springer.com/article/10.1007%2Fs10677-016-9745-2. Note also Lin in https://www.theatlantic.com/technology/archive/2015/10/trolley-problem-history-psychology-morality-driverless-cars/409732/, and Achenbach in https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/.

[5] Goodall, N.J. Ethical Decision Making during Automated Vehicle Crashes. Transportation Research Record. 2014; 2424(1):58-65. doi:10.3141/2424-07.

[6] Achenbach, J., Washington Post (2015). https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/.

[7] Lin, P., Abney, K., and Jenkins, R. (Eds.) (2017). Robot ethics 2.0: From autonomous cars to artificial intelligence. Oxford University Press.

[8] Nyholm, S., and Smids, J. The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem? Ethical Theory and Moral Practice 19, 1275-1289 (2016). The Trolley Problem is a thought experiment that presents two situations that offer similar choices and potential consequences (Foot 1967). See more at https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/.

[9] Roff, H.M., https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles.

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

[11] Mercado et al., 2016. Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A.K., Yang, X.J., and Robert Jr, L.P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. Transportation research part C: emerging technologies, 104, 428-442.

12. Duttaroy, A. 3 X's of Explainable XAI. White paper, A Larsen and Toubro Group Company.

13. Hacker, P., Krestel, R., Grundmann, S., and Naumann, F. (2020). XAI under contract and tort law: legal incentives and technical challenges. Artificial Intelligence and Law, 1-25.

14. The GDPR's fundamental principles are the following: accountability, transparency, and fairness. Transparency is included as a fundamental principle for the protection of personal data included in Article 5(1)(a) of the GDPR. It applies to any processing, regardless of the basis. Accountability is a core theme of the GDPR. The requirement of accountability is to ensure that organizations not only comply with the GDPR, but can demonstrate their compliance. The GDPR states in Article 5 that an organization must take appropriate technical and organizational measures to protect personal data and be able to demonstrate that these have been taken. A "Privacy by Design and Default" approach is required, which means that organizations must take safeguarding measures before starting a personal data processing activity and throughout the life cycle.

15. https://edpb.europa.eu/sites/default/files/files/file1/edpb_letter_out2020_0004_intveldalgorithms_en.pdf.

16. Artificial Intelligence and Data Protection—How the GDPR Regulates AI. Centre for Information Policy Leadership ("CIPL"), March 2020.

17. S. Prevost, (2019). On Data Privacy in Modern Personal Vehicles. 10.1145/3372938.3372940. The European Data Protection Board further describes personal data possibly collected in connecter vehicles, see https://www.researchgate.net/publication/337729562_On_Data_Privacy_in_Modern_Personal_Vehicles

18. Lin, P., Abney, K., and Jenkins, R. (Eds.) (2017). Robot ethics 2.0: From autonomous cars to artificial intelligence. Oxford University Press.

19. Selbst, A., and Powles, J. (2018, January). "Meaningful Information" and the Right to Explanation. In Conference on Fairness, Accountability and Transparency (pp. 48-48). PMLR. And Goodman, B., and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." AI Magazine, 38(3), pp. 50-57.

20. Wachter, S., Mittelstadt, B., and Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law, 7(2), 76-99. Also, Edwards, L., and Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. Duke Law and Technology. Review, 16, 18.

21. Recital 71: "such processing should be subject to suitable safeguards, which should include (...) the right (...) to obtain an explanation of the decision reached after such assessment."

22. The GDPR's predecessor had a similar right, which was similarly limited in scope. Wachter et al. explain that the right of access was limited by trade secrets and intellectual property provisions. See Wachter et al.

23. EDPB Guidelines on Automated Individual Decision-Making, 9.

24. District Court of Amsterdam, March 11, 2021, paragraphs 4.25 and 4.26 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBAMS:2021:1018.

25. Hacker, P., Krestel, R., Grundmann, S., and Naumann, F. (2020). Explainable AI under contract and tort law: legal incentives and technical challenges. Artificial Intelligence and Law, 1-25.

26. NJCM cs/De Staat der Nederlanden; case C-09-550982-HA ZA 18-388 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:865.

27. NJCM cs/De Staat der Nederlanden; paragraph 6.85, case C-09-550982-HA ZA 18-388 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:865.

28. NJCM cs/De Staat der Nederlanden; paragraph 6.87, case C-09-550982-HA ZA 18-388 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:865.

29. Hacker, P., Krestel, R., Grundmann, S., and Naumann, F. (2020). Explainable AI under contract and tort law: legal incentives and technical challenges. Artificial Intelligence and Law, 1-25.

30. 5 U.S.C. § 552a.

31. 29 U.S.C. § 1181 et seq.

32. 15 U.S.C. § 6501 et seq.

33. 15 U.S.C. § 6802 et seq.

34. These rights include the right to: (1) know what consumer personal information is being collected by businesses; (2) know whether their personal information is being sold or disclosed, and to whom; (3) opt out of the sale of their personal information; (4) access their personal information; (5) request that a business delete any personal information about them; and (6) equal service and price (i.e., to not be discriminated against) if they invoke their privacy rights. *See* Cal. Civil Code §§ 1798.100, 1798.105, 1798.110, 1798.115, 1798.120 and 1798.125.

35. Cal. Civil Code §§ 1798.140(e) and 1798.140(t)(1).

36. Cal. Civil Code § 1798.140(q).

37. Cal. Civil Code § 1798.140(c).

38. CPRA, Version 3, Section 21 (adding Cal. Civil Code § 1798.185(a)(16)).

39. CPRA, Version 3, Section 14 (adding Cal. Civil Code § 1798.140(z)).

40. The amended data breach notification obligations under the SHIELD Act went into effect on October 23, 2019.

41. New York Consolidated Laws, General Business Law § 899-BB(2)(a).

42. Hua Ling, "Yao quxin-yuren, AI dei dakai juece heixiang (要取信于人，AI得打开决策黑箱)" via Xinhua Net, http://www.xinhuanet.com/tech/2020-03/25/c_1125763152.htm. Retrieved on May 27, 2021.

43. Personal Information Protection Law, Article 24.

44. Personal Information Protection Law, Article 55.

45. Personal Information Protection Law, Article 24.

46. Normative Guidelines on AI Ethics, Guideline 4.1.c.

47. Normative Guidelines on AI Ethics, Guideline 4.2.d.

48. Normative Guidelines on AI Ethics, Guideline 4.3.c.

49. Normative Guidelines on AI Ethics, Guideline 4.4.c.

50. Turek, M. Defense Advanced Research Projects Agency, Program Information, "Explainable Artificial Intelligence" ("XAI"), https://www.darpa.mil/program/explainable-artificial-intelligence.

51. https://www.elementai.com/news/2019/the-what-of-explainable-ai.

52. Glomsrud et al. (2019). "Trustworthy Versus XAI in Autonomous Vessels."

53. https://www.forbes.com/sites/jimgorzelany/2016/09/16/the-smiling-car-concept-gives-autonomous-autos-a-great-big-emoji/?sh=19fae96e243d.

54. https://kambria.io/blog/artificial-intelligence-the-difference-between-machine-learning-and-deep-learning/.

55. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., and Kagal, L. October 2018). Explaining explanations: An overview of interpretability of machine learning. In the 2018 IEEE 5th International Conference on data science and advanced analytics ("DSAA") (pp. 80-89). IEEE.

56. https://www.elementai.com/news/2019/the-what-of-explainable-ai.

57. https://arxiv.org/pdf/1903.10992.pdf.

58. Tjong Tjin Tai, E. (2017). Aansprakelijkheid voor robots en algoritmes. Nederlands Tijdschrift voor Handelsrecht, 14(3), 123-132.

59. Zollers, F.E., McMullin, A., Hurd, S.N., and Shears, P. (2004). No more soft landings for software: Liability for defects in an industry that has come of age. Santa Clara Computer and High Technology Law Journal, 21, 745.

60. Wolpert, Thomas G. Product Liability and Software Implicated in Personal Injury, 60 DEF. CoUNS. J. 519 (1993).

61. Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics, 2020. https://ec.europa.eu/info/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0_en. Bracketed text inserted by author.

62. European Parliament (2017). https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html?redirect#title1.

63. November 21, 2010, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=63199.

64. Expert Group on Liability for New Technologies (2019), p. 47. Para 22.

65. Expert Group on Liability for New Technologies, p. 47. Para 23.

66. Expert Group on Liability for New Technologies, p. 47. Illustration 14.

67. Expert Group on Liability for New Technologies, p. 47. Illustration 16.

68. Expert Group on Liability for New Technologies, p. 50. The allegation of the burden of proof in case of informational asymmetry is similar for instance in the field of medical operations, where medical practitioners have more information than the victim.

69. Bertolini, A. European Parliament, Artificial Intelligence and Civil Liability (2020), p. 83. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/621926/IPOL_STU(2020)621926_EN.pdf.

70. General Product Safety Directive (Directive 2001/95/EC).

71. The consultation's Final Report of White Paper on Artificial Intelligence (2020), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68462.

72. The consultation's Final Report of White Paper on Artificial Intelligence (2020), p. 15.

73. Zech, H. (April 2021). Liability for AI: public policy considerations. In ERA Forum (Vol. 22, No. 1, pp. 147-158). Springer Berlin Heidelberg.

74. Zech, H. (April 2021). Liability for AI: public policy considerations. In ERA Forum (Vol. 22, No. 1, pp. 147-158). Springer Berlin Heidelberg.

75. OMB M-21-06, Guidance for Regulation of Artificial Intelligence Applications (November 17, 2020) at p. 1, available at https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf.

76. OMB M-21-06, Guidance for Regulation of Artificial Intelligence Applications (Nov. 17, 2020) at p. 5, available at https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf.

77. OMB M-21-06, Guidance for Regulation of Artificial Intelligence Applications (Nov. 17, 2020) at p. 5, available at https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf

78. The Shenzhen Proposal (zhineng wanglian qiche), Article 47(1).

79. Shenzhen Municipal Government's Proposed Regulations on Smart Net-Connected Cars, Article 47(2).

80. Shenzhen Municipal Government's Proposed Regulations on Smart Net-Connected Cars, Article 47(3).

81. Shenzhen Municipal Government's Proposed Regulations on Smart Net-Connected Cars, Article 48.

82. Liu Kai, "Zidong jiashi chehuo, shui lai dan ze? (自动驾驶车祸，谁来担责)" Sina Technology, https://finance.sina.com.cn/tech/2021-04-23/doc-ikmxzfmk8496559.shtml. Retrieved on May 26, 2021.

83. Such issues include: the decision may be affected by biases in data that the AI is trained in; the algorithm may not be advanced enough to tackle

"open-ended questions"; AI is good at ethical calculations but cannot make any "judgement" per se.

84. Liu points out that the keyword in driving AI is "automation," not "autonomy."

85. Yang Lixin, "Zidong jiashi jidongche jiaotong shigu zeren de guize sheji (自动驾驶机动车交通事故责任的规则设计)," Journal of Fujian Normal University (Philosophy and Social Sciences Edition), 2019 vol. 3.

86. Zhang Jihong and Xiao Jianlan, "Zidong jiashi qiche qinquan zeren wenti yanjiu (自动驾驶汽车侵权责任问题研究)," Journal of Shanghai University (Social Sciences Edition), 2019 vol. 1.

87. Curtis G Northcutt and others, 'Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks' [2021] arXiv:210314749 [cs, stat] en M. Veale & F. Zuiderveen Borgesius, Demystifying the Draft EU Artificial Intelligence Act, Computer Law Review International 2021/22(4).

88. Ha, T., Lee, S., and Kim, S. (2018). Designing explainability of an artificial intelligence system. In Proceedings of the Technology, Mind, and Society (pp. 1-1) en Rai, A. (2020). Explainable AI: From black box to glass box. Journal of the Academy of Marketing Science, 48(1), 137-141 bijvoorbeeld.

89. District Court of Amsterdam, March 11, 2021, paragraphs 4.25 and 4.26 https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBAMS:2021:1018.

90. Hacker, P., Krestel, R., Grundmann, S., and Naumann, F. (2020). XAI under contract and tort law: legal incentives and technical challenges. Artificial Intelligence and Law, 1-25.

91. Algorithmic Accountability Act of 2019, H.R. 2231, 116th Congress, 1st Session (2019), available at https://www.congress.gov/116/bills/hr2231/BILLS-116hr2231ih.pdf.

92. The NAII Act, H.R. 6216, 116th Congress, 2nd Session (2020), available at https://www.congress.gov/116/bills/hr6216/BILLS-116hr6216ih.pdf.

93. FTC Business Blog, "Using Artificial Intelligence and Algorithms" (April 8, 2020), available at https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms.

94. FTC Business Blog, "Aiming for truth, fairness, and equity in your company's use of AI" (April 19, 2021), available at https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai.

95. Final Report of the National Security Commission on Artificial Intelligence at p. 15, available at https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

96. Final Report of the National Security Commission on Artificial Intelligence at p. 8, available at https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

[97] Final Report of the National Security Commission on Artificial Intelligence at pp. 148-149, available at https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

[98] Hamon, R., Junklewitz, H. and Sanchez, I. (2020). Robustness and explainability of artificial intelligence. Publications Office of the European Union.

[99] One also could contend that the assertion of other intellectual property rights may hinder the transparency of AI. This likely goes too far. Under many circumstances, general algorithms likely are not covered by patent and copyright law.

[100] https://www.wto.org/english/docs_e/legal_e/27-trips_04d_e.htm.

[101] Meyer, D. (May 25, 2018). AI has a big privacy problem and Europe's new data protection law is about to expose it. Fortune.

[102] Directive (EU) 2016/943 of the European Parliament and of the Council of June 8, 2016, on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure.

[103] https://link.springer.com/article/10.1007/s12027-020-00648-0#Fn25.

[104] https://ec.europa.eu/transparency/expert-groups-register/screen/home?do=groupDetail.groupMeetingDoc&docid=36608.

[105] Caplan, Donovan, Hanson and Matthews, 2018; Miller, 2019.

[106] De Laat, P.B. (2018). Algorithmic decision-making based on machine learning from big data: can transparency restore accountability?. Philosophy & Technology, 31(4), 525-541.

[107] Lipton, Z.C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31-57.

[108] Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Wortman Vaughan, J.W., & Wallach, H. (2021, May). Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-52).

[109] Edwards, L., and Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. Duke Law and Technology Review, 16, 18.

[110] Edwards, L., and Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. Duke Law and Technology Review, 16, 18.

[111] Edwards, L., and Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. Duke Law and Technology Review, 16, 18.

[112] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., and Kagal, L. October 2018). Explaining explanations: An overview of interpretability of

machine learning. In the 2018 IEEE 5th International Conference on data science and advanced analytics (“DSAA”) (pp. 80-89). IEEE.

[113] AI HLEG, 2018, p. 6.