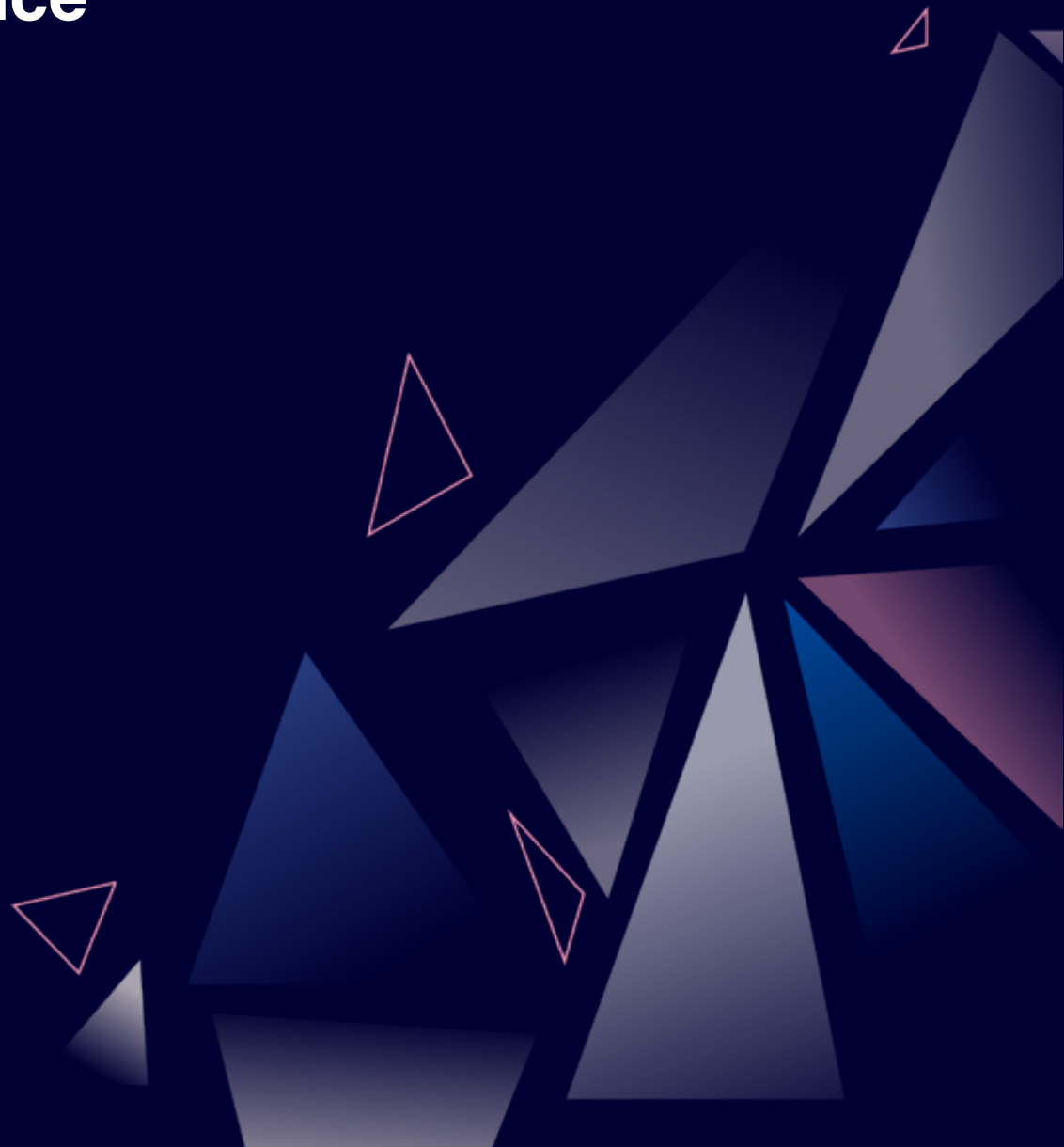



V29 Legal's Contribution to the European Commission's Public Consultation on the White Paper on Artificial Intelligence





V29 Legal is a boutique law firm specializing in international dispute resolution as well as legal matters related to artificial intelligence (AI). In this regard, V29 Legal works with innovators to create trustworthy AI.

V29 Legal submits this contribution to the European Commission's Public Consultation on the White Paper on Artificial Intelligence (*Contribution Paper*) to strengthen the potential of AI within the European Union (EU) while identifying and minimising its risks. The Contribution Paper focuses on the transparency of AI systems as one of the key components of *trustworthy AI*. It explores how a future regulatory framework could spur innovation and, at the same time, address risks in a manner that enables creation of an ecosystem of excellence and trust.

Table of Contents

A. Executive Summary	1
B. Introduction	2
C. Risks related to opacity - a case of explainable AI	3
I. Accountability and traceability	3
II. Human control and risk prevention	4
III. Accuracy and prevention of bias	5
IV. Contestability and action	6
V. Trust in development of AI	6
D. The five dimensions of transparency	7
I. Identification transparency: Is an AI-driven system being used?	7
II. Context transparency: Why does the AI-driven system exist?	7
III. Function transparency: How does the AI-driven system work?	7
IV. Explanation transparency: What is the rationale behind the decision?	8
V. Option transparency: What options do I have?	9
E. Existing legal obligations regarding transparency and explainability	9
I. The General Data Protection Regulation	9
II. Other regulations	10
1. Administrative decisions	10
2. Court decisions	10
3. Private actions	10
4. Liability in contract and tort law	10
F. De lege ferenda: A call for a gradual approach	11
I. Critical sectors	12
II. Risk matrix	13
1. Risk-level (1): low risk - ex-post traceability	13
2. Risk-level (2): moderate risk - basic function monitoring	13
3. Risk-level (3): medium risk - individual decision explanation	14
4. Risk-level (4): elevated risk - pre-market approval	14
5. Risk-level (5): high risk - prohibition of self- or deep-learning algorithms	15
III. Potential trade-offs and technical limits	15
IV. How to implement transparency and explainability	16
1. Implementing transparency	16
2. Implementing explainability	16
G. Conclusion	19

A. Executive Summary

- 1 Acknowledging the increasing importance of artificial intelligence in modern society, and the expected benefits, with this Contribution Paper, V29 Legal invites the European Commission (*EC*) to consider the following key findings when drafting a new framework on AI:
- i. Transparency of AI systems is a key component of *trustworthy AI*. Therefore, a regulatory framework needs to provide for the degree of transparency that creates and sustains trust.
 - ii. A future European regulatory framework needs to address AI-related risks in a differentiating and detailed manner, taking differences between various AI applications and technological possibilities and their respective risks into account.
 - iii. Such a differentiating system needs to be accompanied by a monitoring and enforcement mechanism.
 - iv. In practical terms, this means that:
 - a) fully transparent AI systems must be clearly identifiable as such, state their purpose, indicate their underlying functioning in such a way that independent performance tests can be carried out, offer an *ex post* explanation of individual decisions and contain information on alternative courses of action.
 - b) a new European regulatory framework for AI needs to clearly define which transparency requirements are mandatory, and, based on a risk assessment, classify AI applications into five different risk levels, each with increasing transparency and monitoring obligations.
 - c) explainable AI methods will be a key requirement/tool to explain complex neural networks with self-learning or deep learning capabilities.

B. Introduction

- 2 The futurist Raymond 'Ray' Kurzweil has predicted that AI will outsmart human intelligence by 2029.¹ Whether the timing of the prediction is fully accurate, is not of great importance when it comes to the question of a future regulatory framework for AI. Today, AI is already being developed and deployed in various areas which have the potential to outpace human development. While these technologies bring about tremendous advantages, they also bear great risks.
- 3 Artificially intelligent systems are playing an increasingly central role in our daily lives and are no longer mere decision *aids*, but are increasingly deployed as decision *makers*. In order to preserve a human society where humans continue to make and to bear responsibility for the final decisions, we have to ensure that humans remain in control. This approach is reflected in the European Commission's White Paper on Artificial Intelligence (**White Paper**), which emphasized the 'human centric approach' towards regulation of AI as its core principle.² The European values-based, ethical AI approach considering 'human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities'; and a 'society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail' has been already defined as a 'secret sauce'³ or a 'secret weapon'⁴ of the EU in global AI race.
- 4 In this context, the White Paper further acknowledges Europe's 'strong attachment to values and the rule of law as well as its proven capacity to build safe, reliable and sophisticated products and services from aeronautics to energy, automotive and medical equipment'.⁵ In addition, the White Paper emphasizes the importance of trust into AI systems, as a prerequisite for its uptake.⁶ The White Paper presents various 'policy options to enable a trustworthy and secure development of AI in Europe, in full respect of values and rights of EU citizens'.⁷ With the Public Consultation, the EC aims to give stakeholders the opportunity to state their views on the questions raised and policy options proposed in the White Paper.
- 5 With the defined goals of the Public Consultation in mind, this Contribution Paper suggests that one of the key aspects to retaining human control over AI is the creation of transparent systems. To this end, we first present a brief outline of challenges arising due to the opacity of the so-called black-box systems (**C.**). Then, possible questions with regards to AI systems which should be answered in order to ensure transparency and the necessary degree of transparency are explored (**D.**). In a third step, follows an overview of existing legal obligations regarding transparency and explainability which already exist in some established areas (**E.**). This analysis is then followed by specific proposals for future transparency requirements for AI (**F.**). In addition, proposals for possible translation of these requirements into practice are made (**G.**), including an overview of trade-offs and obstacles which the implementation of such explainability may face (**H.**).

C. Risks related to opacity – a case for explainable AI

- 6 AI-driven systems are typically characterised as so-called black-boxes. This characterization is a result of three interrelated dimensions of algorithmic opacity:
- 7 First of all, and unlike early forms of algorithmic systems, which relied on 'if-then' reasoning, modern machine learning systems can create complex models which can make it difficult or even impossible to identify why and how they generate a particular output.
- 8 Secondly, even systems that utilize algorithms whose underlying operation and logic can be explained (for example, because they follow a decision-tree analysis) may not openly display their reasoning.
- 9 Thirdly, if information about a system is provided, end-users will often not be able to comprehend or assess such information because of the quantity of information and complexity of these systems.⁸
- 10 As a result, five different but interrelated challenges arise. To overcome such challenges, an explainable AI system needs to respect the following principles: accountability and traceability (I.), human control and risk prevention (II.), accuracy and prevention of bias (III.), contestability and action (IV.) and trust in the development of AI (V.).

I. Accountability and traceability

- 11 Non-transparent systems may make it difficult or impossible to control and, thus, to monitor compliance with legal requirements of these systems. This is problematic from a regulatory perspective, as – for instance – state agencies can face difficulties in monitoring and enforcing safety requirements. Likewise, from an end-user perspective, the lack of supervision may lead to a lack of verifiability of causal chains of errors and, therefore, ultimately to liability gaps.
- 12 In addition, an error may not result from the operation of the system itself but be the consequence of a simple human input error. However, even the origin of input errors may be difficult to determine if the system does not allow for traceability.⁹ In this regard, transparency can help to expose a basis for evaluation.¹⁰



Autonomous driving

The difficulties in determining the root cause were raised, for example, in connection with a tragic accident that occurred in March 2018 in Arizona. A driverless Uber-operated Volvo car killed a woman. Subsequently, it was difficult to trace back the chain of events that led to the accident: Was it the built-in sensor and should, therefore, its manufacturer be held liable? Was it a mechanical failure for which Volvo should be liable? Was it the camera? Was it Uber that runs the operating system of these cars?

The lack of transparency of the AI system steering the vehicle made it even more difficult to understand how the system had reacted and therefore whether the AI itself, or individual elements of the self-driving system (the camera, the sensor, safety settings) had set the decisive cause. The investigation also focused on the possible distraction of the accompanying driver. Furthermore, the victim was reportedly under the influence of drugs and alcohol. At the same time, it was unclear whether this state might have affected the predictability of her behaviour or raised questions of contributory negligence. Ultimately, it

turned out that Uber had disabled at least two safety-critical functions, including emergency braking.



Credit risk assessment

One of the application cases of AI in the financial sector are AI-based credit risk assessment methods, be it by eCommerce businesses or by financial institutions. In this context, difficulties with regards to accountability and traceability, may, *inter alia*, arise when a loan is denied.

Payment solutions providers like, for example, PayPal and Klarna, have introduced delay-payment-based solutions, e.g. open invoice (purchase on account) or payment in instalments. After selecting the preferred delay-payment method, the users are requested to enter their full name, date of birth, address,¹¹ and accept the providers' terms and conditions. These user data are then interpreted by the AI system and a decision on the user's financial solvency is made. As the initial date set is entered by the user, it is known to him or her. However, it is non-transparent to the user what other data sets (e.g. training data) the search query is run against. In addition, when the selected payment method is denied, the decision-making criteria, i.e. a reasoning for the refusal, are generally not presented by the provider.

Similarly, financial service providers like, for example, Postbank and ING, are increasingly relying on fully automated instant credit offerings to meet the consumers' needs. Through the use of digitised tools like document-upload, video identification, and eSignatures, loans nowadays can be granted almost instantaneously, with only one to two days between application and pay-out. However, in case of denial, companies tend to use formulaic and standardized language.

This example shows the benefits and downsides of the use of AI: on the one hand, AI may help lenders to measure credit risk, limit default risks and expedite the procedure. At the same time, the affected consumers may not be able to understand why their application is declined and accordingly effectively challenge such decision or improve their scoring, since they are not provided with reasons for the decision. They may also not be aware of the data which have been used. For example, some AI models even rely on the customers' behaviour in social media.¹²

II. Human control and risk prevention

- 13 Apart from general accountability aspects, the opacity of AI-systems may make it even more difficult to ensure human control over these applications. Such lack of control can lead to catastrophic outcomes.



Autopilot

This became apparent in the case of the two Boeing 737 Max crashes on 29 October 2018¹³ and 10 March 2019¹⁴. It seems as if the main problem in the case of the Lion Air crash of October 2018 was that the producer had installed an AI-driven system (the 'Manoeuvring Characteristics Augmentation System'). This system was supposed to ensure that the nose of the aircraft is automatically pushed down if a stall threatens. The final accident investigation report found, amongst other contributing factors, however, that the pilots had neither been informed of its existence nor trained in its operation.¹⁵

Obviously, no pilot should ever be confronted with any such risk. However, if the said system had at least been duly explained to the pilots, the risks might have been recognized and the crashes been prevented. The extreme example shows how important the understanding of intelligent systems and their functioning (and handling) within a superordinate system may be.

III. Accuracy and prevention of bias

- 14 Thirdly, the opacity of self-learning algorithmic systems may help hiding potentially biased or erroneous decisions. In this regard, transparent systems are necessary to protect the inherent good of society.¹⁶



Healthcare sector

An investigation revealed in 2018 that IBM's Watson supercomputer recommended unsafe and incorrect cancer treatments: Internal IBM documents showed that the software was drilled with a small number of hypothetical rather than real patient data. These hypothetical data were not representative and, *inter alia*, reflected some doctor's preferential treatments. The hospitals who used the software did not know that it was trained on such hypotheticals. Furthermore, recommendations were based on the expertise of a few specialists for each cancer type instead of guidelines or evidence. In addition, the internal document revealed that the IBM software-based studies that were advertised as a proof of the system's usefulness, were designed to generate favorable results.¹⁷



Bias in Human Resources systems and financial sector

A similar problem is that non-transparent systems can unintentionally trigger or reinforce discrimination. To name just one of the many examples, Amazon found in 2015 that its recruitment algorithm categorically disadvantaged women. Amazon's model was trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Since these (due to the male-dominated tech sector) were mainly male, the system had learned to discriminate against women and therefore in the following evaluated activities in women's associations or the completion of women's colleges as negative.¹⁸

- 15 If the system had been more transparent – and, in particular, had disclosed its decision-making process, the developers might have realised much earlier that the system had an inherent bias. In addition, the disclosure of the decision-making process would have enabled independent third parties or affected individuals to discover the bias and take action against it. Without this information, they could only speculate or use examples to demonstrate that women were less successful.
- 16 Similarly, even an arguably unbiased initial set of data, e.g. without any reference to gender or race, could lead to a discriminatory decision in a credit risk assessment procedure. For example, based on a default-rate among residents of a certain district, an indirect bias might be developed by the AI system with regard to the creditworthiness. Transparency is, thus, also needed with respect to the data in order to avoid a negative risk assessment, for example with regard to minority groups.

IV. Contestability and action

- 17 Fourth, opaque systems may hinder access to justice if an individual AI-decision does not offer any reasoning. However, transparency is essential to enable legal action.¹⁹ In order to contest a decision in court, for example, each individual needs to be able to rely on evidence.



Credit risk assessment

If a bank refuses to issue a loan to a customer based on the decision of an AI-driven credit scoring system, he or she is unlikely to effectively contest that decision without a possibility to analyse and rely on the reasoning of and criteria behind the AI system's decision. AI-based decisions which do not reveal their reasoning or at least criteria underlying such decisions, may, therefore, jeopardizes each individual's right to effectively contest a certain decision.



Legal system

The need to rely on facts and reasoning applies even more so in situations in which the court itself makes use of an AI-driven system that it cannot assess. This concern came to head, for example, in *State v. Loomis*, a case in which the Supreme Court of Wisconsin on July 13, 2016, based the sentence for a drive by shooting on the result of the so-called COMPAS ('Correctional Offender Management Profiling for Alternative Sanctions') tool. The software was used to predict the probability of the defendant's recidivism.²⁰ The defendant challenged the sentence, arguing that the court's use of the recidivism prediction tool suggested that it violated his due process rights as the court was not able to assess the system's accuracy.²¹ The algorithms used were considered trade secrets and the causal audit process was not clearly known to the judge.²² The court, however, rejected the claim concluding that

'...if used properly, observing the limitations and cautions set forth herein, a circuit court's consideration of a COMPAS risk assessment at sentencing does not violate a defendant's right to due process'.²³

- 18 This case shows that appealing against an automated decision is difficult, if not impossible, if the AI system's functionality and decision-making criteria are not available to the individual affected by the system. Potential discrimination in this specific case was only established later on as a result of an analysis by ProPublica, an independent investigative journalism platform.²⁴

V. Trust in the development of AI

- 19 Lastly, the fact that most complex algorithmic systems are perceived as black boxes may also prevent certain actors from deploying AI for fear of improper or illegal decisions and consequences. A lack of understanding of the system's functionality and, hence, a lack of trust in the system as such, can prevent the usage of the AI-based products and services. Ultimately, insufficient transparency can, therefore, limit related economic growth and competitiveness.
- 20 According to research conducted by the McKinsey Global Institute, the potential of Europe to deliver on AI is large, as it could add up to EUR 2.7 trillion in GDP by 2030. This would result in a cumulative growth of 19 percent or a 1.4 percent compound annual growth through 2030. Europe's ability to capture the full potential of AI varies significantly among sectors and

countries. With the exception of some Scandinavian countries and the United Kingdom, Europe lags behind the United States in its readiness for AI. For instance, only two European companies are in the worldwide digital top 30, and Europe is home to only 10 percent of the world's digital unicorns.²⁵ At the same time, a creation of a solid values- and rules-based regulatory framework within the EU, could spur AI's development and application within the EU and strengthen the EU's global position in this context.

D. The five dimensions of transparency

- 21 In order to address the outlined challenges and to adhere to the laid-out transparency principles, transparent AI systems and assessment methods need to be designed. To this end, any so-called explainable AI needs to answer the following five questions:

VI. Identification transparency: Is an AI-driven system being used?

- 22 Every user needs to know he or she is dealing with AI. This aspect can be described as **identification transparency**. Each AI-application should be easily and clearly identifiable as such for end-users. Individuals should have a right to know whether they are subject to an autonomous decision-mechanism or not.
- 23 For example, within the context of the above-mentioned credit risk assessment in e-commerce, during check-out on a vendor's website, the customer should be able to identify (upon first glance) that certain payment methods like purchase on account or payment in instalments will only be available in case of a positive AI-based decision with regard to the customer's credit scoring.

VII. Context-transparency: Why does the AI-driven system exist?

- 24 This question deals with the so-called **context-transparency**. The key here is that operators should disclose the intended purpose of using an AI-system.
- 25 To stay with the e-commerce example, the customer should be made aware that credit scoring is mainly employed to identify and/or estimate a customer's default probability. In addition, it is used for identification (identity check) of the customer, general fraud recognition (fraud check), as well as the above-mentioned creditworthiness (credit check).

VIII. Function-transparency: How does the AI-driven system work?

- 26 The underlying dimension of this question is the so-called **function-transparency**: Function-transparency is needed to allow a general and meaningful oversight over the system. Meaningful oversight requires AI operators to disclose the basic features of the system's underlying functioning. This is needed to allow independent third parties to conduct independent testing of the technology, if so required. Independent testing in its turn should include accuracy and unfair performance tests. Depending on the use case, the results of such tests as well as the information regarding the system's error rate should also be made public, obviously without revealing business secrets or any other sensitive data.

- 27 To implement such a requirement, a provider of online payment solutions, for example, would have to present a general description of the underlying technique, to an independent third-party auditor or a supervisory authority. The company would need to disclose, at a minimum, basic features of the AI-system with a particular emphasis on the provenance and quality of the training data the system has been provided with. In addition, the provider would need to disclose who they are collaborating with (e.g. Risk Management Consultants), and in what ways and to which degree the AI-applications interact and/or intersect.

IX. Explanation-transparency: What is the rationale behind the decision?

- 28 This question raises the issue of **explanation-transparency**: Explanations expose information about specific individual decisions without necessarily exposing the precise mechanics of the decision-making process.²⁶ Viewing transparency as 'explaining the steps of the algorithm' is unlikely to lead to an informative outcome.²⁷
- 29 At the same time, the demand for full disclosure of all technical functions of an algorithm could also prove counterproductive. Similar regulatory approaches, such as the duty to consent to the processing of personal data under the GDPR, have demonstrated the risk of information overload: If individuals are being provided with too much information about a transaction they cannot comprehend, the risk of creating even more opacity arises.
- 30 Therefore, 'explanation' should be defined to mean human-interpretable information about the logic by which a decision-maker took a particular set of inputs and reached a particular conclusion.²⁸ It, thus, also serves as a tool to ensure that the individual decision can be contested by the individual affected. As seen in the examples above, individuals are not able to challenge a decision based on an AI driven system if no usable evidence can be extracted from it.
- 31 In order to reach this goal, the following information should be provided about the basic features of the **individual decision** in question, namely:
- The criteria used,
 - their weightings and
 - the training data of the self-learning algorithm.
- 32 In the context of online purchases, after being presented with the aforementioned information, the customer should be able to provide the necessary data to proceed and initiate the AI-application's credit analysis. The company then would need to explain the rationale of the decision. Usually, the customer's data would include the first and last name, her or his date of birth, and address. In some cases, a phone number may be required as well. If, after processing, the AI-application concludes that a specific payment method cannot be offered, the customer should be informed of that decision, and, at a minimum, be provided with an individualised explanation as to the reasons why his request was declined. The explanation should not be a technical one, but, at a minimum, clarify which reasoning was flagged by the AI-application. For example, the customer has recently moved and entered a new address. This caused the fraud-indicator to spike and decline the selected payment method in this specific case.

X. Option-transparency: What options do I have?

- 33 This question sheds light on the so-called **option-transparency**: Individuals should be informed in clear and understandable language whether or not the solutions offered by the AI tool are binding and if they have alternative options.
- 34 Generally, once the selected payment method on an e-commerce platform has been declined, the decision is final and may not be contested by the customer. In some cases, when a payment solution is not linked to a specific transaction, for instance, when applying to a service like PayPal Credit, there is an option to re-apply after a certain period of time has passed. The payment solution provider would need to implement a dispute resolution channel, similar to the Resolution Centre offered by PayPal, that extends the possibility for the customer to appeal the decision made by the payment solution provider, therefore, ensuring that an individual and analogue credit analysis can be enforced.

E. *De lege lata*: Existing legal obligations regarding transparency and explainability

- 35 At present, specific requirements regarding the transparency or explainability of algorithmic decisions exist mainly in the area of data protection law. However, the following analysis will show that there is a need for the explainability of decisions in other areas of law as well. This necessarily means that when AI systems are used in the context of such decisions, they must be explainable in order to meet these general requirements. This shows that the demands for transparent systems are already deeply rooted in most legal systems. They should, therefore, only be specified in an overarching EU legal framework on AI for clarification and uniformity purposes.

I. The General Data Protection Regulation

- 36 In 2016, the EU adopted the General Data Protection Regulation (**GDPR**), which took effect on May 25, 2018, and replaced the 1995 Data Protection Directive.²⁹ The GDPR's discussion of automated decisions is contained in Articles 22, 13(2)(f), 14(2)(g), and 15(1)(h) of the GDPR.
- 37 The provisions demand that the following information be available to data subjects: '*the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject*'. Additionally, the non-binding Recital 71 which corresponds to Article 22 GDPR explicitly grants the individual to 'obtain an explanation of the decision reached'.
- 38 Since adoption of the GDPR, scholars have debated whether these requirements amount to a '*right to explanation*.' Some argue that if not the text of the norms themselves, at least the combination of the right to meaningful information about the logic involved and the right to contest a particular decision implies a right of explanation.³⁰ Others argue that an explanation of individual decisions is not provided for, but is in any case unnecessary and merely distracts the discussion.³¹
- 39 This question can be left open for the purposes of this Contribution Paper, since it is in any event clear that the data subject has both, a right to be informed about the overall functioning of the system and, at least to a certain extent, about the individual decision itself.

II. Other regulations

- 40 The duty to explain an outcome of an automated decision-making system has been mostly debated in the ambit of data protection. In the following, however, we will demonstrate that explanation, or – more specifically – the duty to state reasons for certain decisions is a far more widespread concept and already amounts to a legal requirement across jurisdictions and sectors. In the following, we will highlight certain areas where this observation leads to an imperative demand to create explainable AI-systems.

1. Administrative decisions

- 41 States and state-owned actors are addressee of the duty to explain their decisions which typically, in a democratic state, requires a state to provide reasoning for its acts, i.e. for administrative decisions. This concept is embodied in Article 41(2) of the Charter of the Fundamental Rights of the European Union (*Charter*). It has further been specified in national legislations across all jurisdictions. For instance, under the German Administrative Procedure Act, an agency that enacts a decision must provide an explanation that include the 'chief material and legal grounds' for the decision.³² France has recently enacted the Digital Republic Act, which creates a right³³ for subjects of algorithmic decision-making by public entities to receive an explanation of the parameters (including specification of their weighting) used as a basis for the decision.³⁴

2. Court decisions

- 42 Court decisions usually have to be explained, e.g. be accompanied by a reasoning of a judge. In the United Kingdom, for example, it is a principle of common law that a judgment must be reasoned; if a judgment is not sufficiently explained, it can be overturned by a higher court.³⁵ In France³⁶ and Germany³⁷, the respective civil codes explicitly stipulate that all judgments must be justified. Again, the failure to make a declaration may lead to the judgment being set aside.³⁸

3. Private actions

- 43 In certain areas, even private actors are required to explain their acts. In the U.S., for instance, the Fair Credit Reporting Act requires consumer reporting agencies to provide, with every request for a credit score, a list of the key factors that negatively influenced the consumer's score. This provision permits consumers to contest their credit scores, thereby adding a layer of accountability to the system.³⁹ Similarly, the Equal Credit Opportunity Act requires a creditor to give a statement of specific reasons if he takes an adverse decision.⁴⁰

4. Liability in contract and tort law

- 44 Furthermore, and most crucial for the development of explainable AI, explainability is an important legal category in contract and tort law. In fact, contract and tort law may impose legal requirements to use explainable machine learning models.⁴¹
- 45 For example, an explanation for a particular choice may provide evidence of whether the defendant acted knowingly, recklessly, negligently or innocently - all of which can have legal

significance. The precise amount of evidence required to compel an explanation varies with the governing law.⁴²

F. *De lege ferenda: A call for a gradual approach*

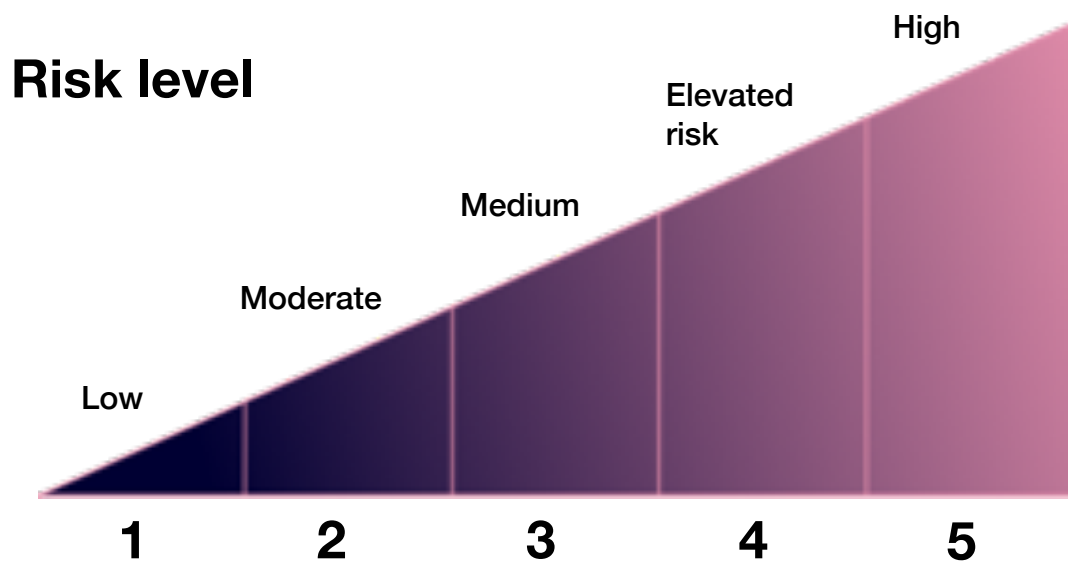
- 46 In the White Paper, the Commission stated that a clear European regulatory framework for trustworthy AI is needed to avoid fragmentation in the internal market, which would undermine the objectives of trust, legal certainty and business uptake.⁴³ With regard to transparency, the White Paper indicates that a future AI framework could include the requirement to provide information on the purpose of the system, the conditions under which it is expected to function and the expected level of accuracy.⁴⁴ Furthermore, the White Paper states that individuals should be clearly informed when an AI system is being used.
- 47 The White Paper already includes references to the requirements that we have summarized under the terms of identification-transparency and context-transparency (and probably also function-transparency). The White Paper's requirements currently, however, do not include any right to obtain an explanation of a decision (explanation-transparency) or to be informed about possible options with regard to the outcome (option-transparency). Furthermore, it does not clearly lay down what requirements about the system's functions are to be revealed. As seen in our analysis above, however, clear rules with regard to the specific information about a system's decision-making process are needed in order to challenge it effectively.
- 48 Similarly, the Ethics Guidelines for Trustworthy Artificial Intelligence prepared by the High-Level Expert Group on Artificial Intelligence (*Guidelines*) introduce transparency as one of their key principles. The Guidelines require in particular, that especially fundamental rights sensitive AI-system need to be explicit and open about choices and decisions concerning data sources, development processes and stakeholders. This implies an element of explainability, without explicitly naming it as such.⁴⁵ The guidelines, however, do not lay down whether non-explainable AI is unlawful or simply unethical.
- 49 Given the importance of the different dimensions of transparency, each of which serves a different purpose and is necessary to enforce further rights, as outlined above, we invite the Commission to develop an EU legal framework for AI that clearly sets out the transparency requirements necessary for each AI application. Otherwise, there is a risk that transparency will become a well-intentioned mute formula. In addition, not every AI application has to meet the same transparency requirements, rather a gradual approach addressing the specifics of AI systems and related risks is necessary.
- 50 At the moment, the White Paper addresses high risk and low risk AI application cases. In view of the complexity and variety of AI application cases and related risk, we suggest developing a more differentiating risk matrix. Such risk matrix should follow a gradual approach and – following the suggestion of the German Data Ethics Commission – divide applications into at least five different risk levels.⁴⁶ Every risk category would, in turn, lead to a different set and degree of legal requirements.⁴⁷
- 51 When assessing potential risks, both - the affected sector as well as individual aspects of the specific AI application - should be taken into account.
- 52 In the following, sectors that are of utmost importance for the systemic functioning of a state will be identified (I.). In the subsequent section, criteria for a concrete risk assessment of an individual application will be developed (II.).

I. Critical sectors

- 53 Wrongful AI applications affecting the high-risk sectors identified in the White Paper, namely healthcare, transport, energy and parts of the public sector (such as asylum, migration, border controls and judiciary, social security and employment services) might create the main risks to society and the functioning of society as a whole. However, other sectors, such as media, would certainly need to be taken into consideration in this context as well. All these sectors need to be categorized as particularly risk-inherent to ensure that basic needs of society can be fulfilled – with or without the use of AI.
- 54 The White Paper currently lacks clear guidance on what constitutes a concrete critical application. We would suggest two major factors should be taken into account, namely the *likelihood of the occurrence of an identified risk* (i) and the *severity of a potential damage* (ii).
- 55 When it comes to the likelihood of the occurrence of an identified risk, the classic risk-assessment regarding the causal link between an action and a likely consequence must be evaluated. Whether the end user has the possibility to re-evaluate and, thus, mitigate the automated decision also plays a decisive role.
- 56 When assessing the severity of a potential damage, both its nature and its likely extent, need to be considered.
- 57 As regards the type of damage, criteria such as whether the envisaged damage would be reversible or irreversible need to be included in the assessment. Another crucial aspect is the type of the affected fundamental right: In this regard, potential physical harm and, thus, a violation of the right to physical integrity (as enshrined in Art. 3 of the Charter) need to be placed at the highest level.
- 58 With regard to the extent of the potential damage, criteria such as the number of individuals affected, the impact on other fundamental rights, the circumstances, frequency and duration of the adverse effect need to be taken into consideration. For example, a threat to the energy or water supply of a metropolitan area or a breakdown of global supply chains to cyber-attacks or a system breakdown would be major threats to society as a whole.

II. Risk matrix

- 59 The classification of a specific application into one of the risk categories should lead to application of different legal requirements. These could be based on the following risk matrix:



1. Risk-level (1): low risk – *ex post* traceability

- 60 At the lowest level of intervention, only *ex post* transparency obligations may be necessary. This would mean that no permanent control processes would need to be installed. However, upon request, the operator of the system should be able to meet at least the requirements of context and system transparency. In concrete terms, the operator must provide the individual (upon request) with information on the purpose of the system and on the general functioning of the system. *Ex post* transparency obligations upon request would enable explainability, and thus, ultimately, that individuals are able to challenge a particular outcome. AI systems used to suggest products to consumers or to display them in a certain order on social networks, for example, could fall under this category.

2. Risk-level (2): moderate risk – basic function monitoring

- 61 At a second risk level, a general monitoring of the system would be required. In practical terms, the system's operator would be obliged to indicate the system's basic functionality i.e. the quality measures and the learning process of the system and disclose how the system relates to the ultimate decision, i.e. the degree to which the decision is influenced or based on the system's output. The monitoring could be performed by external third-party auditors based on the pre-defined criteria.
- 62 With regard to end-users, the system should be clearly identifiable as AI-driven and thus ensure identification-transparency. Additionally, the purpose and, accordingly, context of the system should be made publicly available, for instance, in an FAQ section of the operator's website.

- 63 Dynamic or personalized pricing techniques could fall under this category. The Uber dynamic pricing model, for instance, matches fares to a number of variables such as time and distance of your route, traffic and the current rider-to-driver demand.⁴⁸ It is a system that could potentially be applied in a discriminating manner, but it leaves the user sufficient room for discretion so as not to trigger any particular fundamental rights relevance.

3. Risk-level (3): medium risk – individual decision explanation

- 64 At a third level, comprehensive transparency obligations may be legally required in addition to monitoring. This would mean that apart from disclosing the purpose and basic functionality, the operator should reveal how a particular decision is reached, at least to certified third parties or government agencies. Operators would need to display all relevant factors and criteria used as a basis for the automated decision, as well as the source and training of data, but without disclosing inner technical functions of the system.
- 65 Third-party monitoring may be accompanied by governmental information and inspection rights. Recording requirements may also apply. The regulations of the German Securities Trading Act (*Wertpapierhandelsgesetz, WpHG*) on algorithmic trading with financial instruments could serve as a regulatory model (Section 6 (4) WpHG).
- 66 Furthermore, the operator would be required to name a responsible person within the company to perform risk management tasks. This person could be held responsible for failures of the systems vis-à-vis third parties.
- 67 As regards the use of AI by the government, purely informative tools used within the administration, such as the so-called ‘*Bobbi*’ chatbot, used by the administration of the municipality of Berlin⁴⁹, could fall into this category. This is because AI systems that are adopted by state actors, even if they are of a purely informative character, should always be subject to increased requirements, as they tend to have a higher fundamental rights relevance. Regarding private operators, AI applications that produce a fully automated decision and that may have an effect on individual’s fundamental rights but are not as critical as to lead to physical harm or to widespread outcomes may fall under this category.

4. Risk-level (4): elevated risk – pre-market approval

- 68 At the next level, a full explanation of the system would be required. This would particularly mean that AI-systems with a learning component may only use explainable methods of machine learning. Such explanation could be given, for example, by means of explanation or decryption algorithms. If business or trade secrets were to hinder the full disclosure of an algorithm’s logic, the business entity would be required at least to disclose them to an authorized supervisory body or agency.
- 69 However, individuals should at the same time be given certain access rights similar to the one foreseen under Art. 15 GDPR. The individual should, based on the information accessed, be able to decide on their own whether he or she wants to take action against the system’s outcome. This would include information not endowed with any significant confidentiality character such as meta-information on training data (descriptive statistics) and quality criteria such as performance metrics.

- 70 Furthermore, preventive admissions procedures in the form of pre-market quality controls could be required to ensure compliance with relevant law prior to its use.
- 71 In addition, ongoing dynamic operator obligations would need to be installed. These would make operators responsible for the results of decisions and the procedural correctness of the system even after its admission to the market.
- 72 Such requirements could apply to AI used within the judiciary and law enforcement or for the examination and allocation of benefits. Regarding private parties, due to the risk to health and life, the need for market authorisation procedures may further apply to applications in areas, such as the health and transport sector.

5. Risk-level (5): high risk – prohibition of self- or deep-learning algorithms

- 73 The highest level of intervention would apply only to exceptional applications which impose systematic and tremendous risks in terms of impact, such as automated lethal weapons. In addition to the previously listed obligations, we would suggest considering limiting applications to non-learning AI systems, i.e. those based on linear regression, or whether other reliable, safe limitations could be found to ensure full oversight and human control of the application.

III. Potential trade-offs and technical limits

- 74 In many cases, the more effective the AI, the harder it will be to explain its decisions in terms humans can understand. Indeed, many AI experts posit an inherent technical trade-off between accuracy and explanation.⁵⁰
- 75 Furthermore, while AI systems can be made explainable, this may result in a trade-off between cost and interpretability. If every step must be documented and explained, the process becomes slower and may be more expensive.⁵¹
- 76 Thirdly, trade secrets may hinder 'opening the black-box'. In this regard, the AI Now Institute argues that AI companies should waive trade secrecy and other legal claims that inhibit full auditing and understanding of their software, because such trade secrecy contributes to the black box effect and makes it hard to assess bias, contest decisions or remedy errors.⁵²
- 77 Mindful of potential trade-offs and the requirements to balance various interests, we suggest that it might not be necessary to establish a compulsory waiver of business secrecy. Instead, as will be shown in the next section and exemplified in the use case on explainable AI (XAI) in fintech, it might be sufficient to introduce various degrees of explainability requirements and resort to XAI technology - which can explain algorithmic systems without necessarily exposing any trade secrets.

IV. How to implement transparency and explainability

78 Of the five transparency dimensions outlined above, the prerequisites can be specified in advance of an application under the categories of identification-transparency, context-transparency, function-transparency and option-transparency, whereas explanation-transparency is only given *ex post*. The first four dimensions are simpler to implement in practice, and the approach of privacy by design can be followed here (1.). It is more difficult, however, to explain the system afterwards, especially if neural, self-learning systems are involved. Therefore, more sophisticated approaches are required here, which are explained below under the keyword explainability (2.).

1. Implementing transparency

79 To implement the above-mentioned identified transparency requirements, the approach of '*regulation by and in design*' could be adopted. The rationale behind '*regulation by design*' is that relevant norms are embedded in the technology itself.⁵³ The concept is inspired by Article 25(1) GDPR which requires controllers by default to process only those data that are strictly relevant for each specific purpose. This idea was further developed and is now recognized under the term of '*ethics by or in design*', meaning that also other relevant requirements may be incorporated into the system itself.⁵⁴ When incorporating the above-mentioned criteria into the architecture of an AI system by default, transparency of the system would be enhanced as these would be clearly identifiable and traceable in the system.

2. Implementing explainability

80 Below we suggest to models to implement explainability of AI: Explainable AI (a.) and blockchain technology (b.).

a. Explainable AI

81 XAI is a concept based on the idea that algorithms provide explanations of their own decisions.⁵⁵ It goes back to the '*Explainable AI*' initiative that was launched in 2016 by the US Defense Advanced Research Projects Agency. This initiative of several organizations and companies aims at developing ways to decode deep-learning algorithms.

82 One way to gain explainability in AI systems is to use machine learning algorithms that are inherently explainable. For example, simpler forms of machine learning such as decision trees, can provide the visibility needed for critical AI systems without sacrificing too much performance or accuracy.⁵⁶

83 As far as neural networks are concerned, different XAI models may be applied for different aspects of explanation.

(1) Exogenous models

84 Exogenous explanation models, for instance, attempt to provide explanation from the outside of the system. As part of these, local-interpretability models provide the subject of a recommendation or decision with information about the characteristics of individuals who received similar decisions.

- 85 Another exogenous method is counterfactuals. Counterfactuals can help individuals understand which factors may have most affected the algorithm's outcome.⁵⁷ This could be used, for instance, in credit score decisions, as in these situations individuals may be interested to know which factor was decisive for the system's outcome.
- 86 On the positive side, counterfactual explanation presents one of the least invasive forms of explanation and can therefore bridge the gap between the interests of data subjects and data controllers that otherwise acts as a barrier to a legally binding right to explanation.⁵⁸ However, the counterfactual explanation has also been criticized as it may mislead by suggesting that some factors have an importance that they do not have.⁵⁹ They achieve a reduction of complexity even in cases where complexity cannot be avoided in an accurate representation of the system. Counterfactual explanations suggest that complex decisions are explained by the causal role of a limited number of features. This can be problematic when, for example, there are, in fact, many features playing an equivalently important, or near equivalently important role.⁶⁰



Credit risk assessment

Such an exogenous model has been introduced by Firamis, a B2B FinTech company that is funded by the EU Research and Innovation programme Horizon 2020.⁶¹ The model follows an agnostic approach and aims at identifying the decision-making criteria of an AI system and variable importance. In addition, the model enables a visualisation of the results on a dashboard.⁶²

It can be applied to various AI systems, e.g. for credit risk assessment when loan decisions are made based on credit scoring platforms. In this particular use case, the model would identify the logic behind a credit risk assessment with a possibility to reveal the decision-making criteria, so that the actual score could be explained, and understood.⁶³

Thus, the model can be of great assistance, when minimizing the risks related to opacity of AI systems. Firstly, it enables the possibility to monitor the decision-making process to ensure accountability and traceability. Secondly, it helps to expose potentially biased or erroneous decisions. Thirdly, the individuals affected by the decision would have the opportunity to not only understand, but also to contest and take further action.

(2) Decomposing explanation models

- 87 Instead of revealing the source code of the algorithms, so-called surrogate models create a second algorithm alongside the original. Such models are able to analyse featured input and output but are not capable to reveal how these are weighed in the system.⁶⁴ Such a system may be suitable for product liability cases, for instance, in order to reconstruct the decision of a driverless car.⁶⁵

b. Blockchain technology

- 88 Another potential means to enhance transparency of AI technology is to make use of the so-called blockchain technology. Blockchain technology is a decentralized, distributed ledger that records the provenance of a digital asset.⁶⁶ Through the use of blockchain technology, immutable records of all the data, variables, and processes are available. The audit trail could be used as evidence and thus help to challenge a decision in court.
- 89 Thus, blockchain technology could make a major contribution to creating transparency of AI. So far, however, this is only cautiously proposed as a solution. One reason is that blockchain technology requires high energy input. Yet another problem arises with regard to the GDPR: Due to the decentralized structure and mode of operation of blockchain, it is not compatible with the primacy of the GDPR, according to which each data processing must be able to appoint a responsible data controller. Furthermore, the fact that transactions in the blockchain can hardly be changed and are, therefore, deemed immutable to hacking is difficult to reconcile with the right to be forgotten, codified in Art. 17 of the GDPR. According to a study by the European Parliament⁶⁷, however, the identified tensions are primarily a result of a lack of certainty on how specific concepts of the GDPR should be interpreted.
- 90 Given the outlined legal insecurities of blockchain in relation to the GDPR, the V29 Legal invites the Commission to provide further regulatory guidance to reconcile these conflicting regimes.
- 91 In summary, different approaches of XAI and other methods can play an important role to achieve AI's explainability. However, mathematical explainability of systems alone must not become a panacea. For example, from the point of view of end users, the specification of formulas alone cannot provide the necessary explanation. Information about the purpose of the AI system and the criteria used can be of greater practical use. At the same time, it would also be too brief to declare XAI unnecessary, as is sometimes done in the literature.⁶⁸ As our analysis has shown, the explainability of decisions is also presupposed for courts or state actors or plays a decisive role in the context of questions of evidence or liability, in which technical details of the system used in each case must be able to be presented for evidence purposes. For this reason, the requirement of explainability always depends on the respective person concerned and his or her situation. In order to always enable the defence against a fully automated decision, the traceability of the systems is, therefore, key. This should be taken into account in future regulation at the EU level.

G. Conclusion

- 92 Transparency of AI systems is one of the key components of *trustworthy AI*. In order to ensure the successful development and deployment of AI systems within the framework of a European values-based system and the rule of law, it is decisive to strike a balance between the specifics of the AI technology and the requirements for the system's transparency and explainability.
- 93 This Contribution Paper suggests that this task can be accomplished by considering different AI-applications and technologies and addressing AI-related risks in a more differentiating and detailed manner. Transparent AI systems are a key requirement to ensure accountability and traceability, human control and risk prevention, accuracy and non-bias, contestability and action and to unlock the full potential and trust in this technology. The Contribution Paper suggests that fully transparent AI systems must be clearly identifiable as such, state their purpose, indicate their underlying functioning in such a way that independent performance tests can be carried out, offer an *ex post* explanation of individual decisions and contain information on alternative courses of action.
- 94 A new European AI framework should build on these transparency requirements and, based on a risk assessment, classify AI applications into five different risk levels, each with increasing transparency and monitoring obligations:
- Risk-level (1): low risk – *ex post* traceability;
 - Risk-level (2): moderate risk – basic function monitoring;
 - Risk-level (3): medium risk – individual decision explanation;
 - Risk-level (4): elevated risk – pre-market approval;
 - Risk-level (5): high risk – prohibition of self- or deep-learning algorithms.
- 95 Finally, in order to reach the required explainability and traceability, different explainability methods, such as XAI and blockchain technology, can be useful to explain AI systems based on complex neural networks with self-learning or deep learning capabilities. Certain methods could, in the long term, be identified by specific labels or standardised through certification at EU level.

Endnotes

- ¹ Loria, Kevin, Kurzweil: Human-level AI is coming by 2029, Business Insider, December 29, 2014, available at: <https://www.businessinsider.com/ray-kurzweil-thinks-well-have-human-level-ai-by-2029-2014-12?r=DE&IR=T> [accessed on June 08, 2020].
- ² European Commission, White Paper: On Artificial Intelligence - A European approach to excellence and trust, February 19, 2020, available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf, p. 3 [accessed on June 12, 2020].
- ³ Renda, Andrea, Artificial Intelligence, Ethics, governance, and policy changes, Report of a CEPS Task Force, Centre for European Policy Studies (CEPS) Brussels, February 2019, available at: https://www.ceps.eu/wp-content/uploads/2019/02/AL_TFR.pdf, p. 45 [accessed on June 03, 2020].
- ⁴ Delcker, Janosch, In global AI race, Europe pins hopes on ethics, April 25, 2018, available at: <https://www.politico.eu/article/europe-commission-andrus-ansip-hopes-ethical-approach-will-be-its-edge-in-global-ai-artificial-intelligence-race/> [accessed on June 12, 2020].
- ⁵ Cf fn. 2.
- ⁶ Ibid.
- ⁷ Ibid.
- ⁸ Yeung, Karen, Responsibility and AI, Council of Europe Study, September 2019, available at: <https://rm.coe.int/responsability-and-ai-en/168097d9c5>, p. 21 [accessed on May 12, 2020].
- ⁹ Tobey, Danny, Explainability: Where AI and liability meet, February 25, 2019, available at: <https://www.dlapiper.com/de/germany/insights/publications/2019/02/explainability-where-ai-and-liability-meet/> [accessed on June 03, 2020].
- ¹⁰ Selbst, Andrew D./Barocas, Solon, The Intuitive Appeal of Explainable Machines, Fordham Law Review 87 (03), December 2018, available at: <http://dx.doi.org/10.2139/ssrn.3126971>, p. 1122 [accessed on June 03, 2020].
- ¹¹ Cf PayPal, Information about the PayPal identity check and data exchange with credit agencies (credit check), December 18, 2018, available at: https://www.paypal.com/de/webapps/mpp/ua/creditchk?locale.x=en_DE [accessed on June 12, 2020].
- ¹² U.S. Department of the Treasury, Opportunities and Challenges in Online Marketplace Lending, May 10, 2016, available at: https://www.treasury.gov/connect/blog/Documents/Opportunities_and_Challenges_in_Online_Marketplace_Lending_white_paper.pdf, p. 20 [accessed on June 08].
- ¹³ Aircraft Accident Investigation Report, available at: http://knkt.dephub.go.id/knkt/ntsc_home/ntsc.htm [accessed June 08, 2020].
- ¹⁴ Aircraft Accident Investigation Bureau Interim Report, available at: https://reports.aviation-safety.net/2019/20190310-0_B38M_ET-AVJ_Interim.pdf [accessed June 08, 2020].
- ¹⁵ Cf fn. 13., p. 215; Langewiesche, William, What really brought down the Boeing 737 Max?, Malfunctions caused two deadly crashes. But an industry that puts unprepared pilots in the cockpit is just as guilty., New York Times Magazine, September 18, 2019, available at: <https://www.nytimes.com/2019/09/18/magazine/boeing-737-max-crashes.html> [accessed on June 03, 2020].
- ¹⁶ Cf fn. 10, p. 1118.
- ¹⁷ Ross, Casey/Swetlitz, Ike, IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show, STAT+, July 25, 2018, available at: <https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf> [accessed on June 03, 2020].
- ¹⁸ Dastin, Jeffrey, Amazon scraps secret AI recruiting tool that showed bias against women, October 10, 2018, Reuters, available at: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> [accessed on June 08, 2020].
- ¹⁹ Cf fn. 10, p. 1120.
- ²⁰ State v. Loomis, Supreme Court of Wisconsin, July 13, 2016, available at: <https://casetext.com/case/state-v-loomis-22>, para 8 [accessed on June 8, 2020].
- ²¹ Deeks, Ashley, The Judicial Demand for Explainable Artificial Intelligence, Columbia Law Review 119 (7), 2019, available at: www.jstor.org/stable/26810851, p. 1844 [accessed on June 03, 2020].
- ²² Ibid.
- ²³ Cf fn. 20.
- ²⁴ Angwin, Julia et al. Machine Bias, May 23, 2016, available at: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [accessed on June 07, 2020].
- ²⁵ McKinsey Global Institute, Notes from the AI Frontier Tackling Europe's Gap in Digital and AI, discussion paper, February 2019, available at: <https://www.mckinsey.com/~media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Tackling%20Eu>

- [ropes%20gap%20in%20digital%20and%20AI/MGI-Tackling-Europes-gap-in-digital-and-AI-Feb-2019-vF.ashx](#)>, pp. 2, 29 [accessed on June 09, 2020].
- 26 Doshi-Velez, Finale et. al, Accountability of AI Under the Law: The Role of Explanation, available at: <https://arxiv.org/pdf/1711.01134.pdf>>, p. 2 [accessed on June 03, 2020].
- 27 European Parliamentary Research Service (EPRS), A governance framework for algorithmic accountability and transparency, April 2019, available at: [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU\(2019\)624262_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf)>, p. 31 [accessed on June 03, 2020].
- 28 Cf fn. 26, p. 4.
- 29 Cf Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), April 27, 2016, available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>> [accessed on June 08, 2020].
- 30 Brkan, Maja, Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond, August 1, 2017, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3124901>, p. 15; Malgieri Gianclaudio/Comandé, Giovanni, Why a Right to Legibility of Automated Decision- Making Exists in the General Data Protection Regulation, International Data Privacy Law, 7 (3), 2017, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3088976>, pp. 243, 245, 250; Mendoza, Isak/Bygrave, Lee A., The Right Not to Be Subject to Automated Decisions Based on Profiling, May 08, 2017, in: Synodinou, Tatiani/Jougleux, Philippe/Markou, Christiana/Prastitou, Thalia (eds.), EU Internet Law: Regulation and Enforcement, Springer, 2017, University of Oslo Faculty of Law Research Paper No. 2017-20, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2964855>; Selbst, Andrew D./Powles, Julia, Meaningful information and the right to explanation, International Data Privacy Law 7 (4), November 2017, available at: <https://academic.oup.com/idpl/article/7/4/233/4762325>> [accessed on June 03, 2020].
- 31 Wachter, Sandra et al., Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, Harvard Journal of Law and Technology, 31 (2), 2018, available at: <https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf>>, p. 841, (arguing that a legal right to explanations of automated decisions does not exist); Edwards, Lilian/Veale, Michael, 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), 2017, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2972855>, p. 44 [accessed on June 03, 2020].
- 32 Sec. 39 para (1) of the German Administrative Procedure Act (Verwaltungsverfahrensgesetz), translation available at: <https://germanlawarchive.iuscomp.org/?p=289>> [accessed on June 10, 2020].
- 33 Art. 4 Law 2016-1321 of Oct. 7, 2016 for a Digital Republic (Loi 2016-1321 du 7 octobre 2016 pour une République numérique) available at: <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id>> [accessed on June 10, 2020].
- 34 Cf fn. 26, p. 11.
- 35 Cf fn. 26, p. 8.
- 36 Art. 455 of the Civil Procedure Code (Code de Procédure Civile), available at: <https://www.legifrance.gouv.fr/Traductions/en-English/Legifrance-translations>> [accessed on June 10, 2020].
- 37 Sec. 313 Para (1) No. 6 of the German Code of Civil Procedure (Zivilprozessordnung), available at: https://www.gesetze-im-internet.de/englisch_zpo/englisch_zpo.pdf> [accessed on June 10, 2020].
- 38 Cf fn. 26, p. 9.
- 39 Ibid.
- 40 Cf fn. 10, p. 1101.
- 41 Hacker, Philipp/Krestel, Ralf/Grundmann, Stefan et al., Explainable AI under contract and tort law: legal incentives and technical challenges. Artificial Intelligence Law, 2020, available at: <https://doi.org/10.1007/s10506-020-09260-6>>, p. 2 [accessed on June 03, 2020].
- 42 Cf fn. 26, p. 11.
- 43 Cf fn. 2, p. 10.
- 44 Cf fn. 2, p. 20.
- 45 Cf fn. 4, p. 67.
- 46 Data Ethics Commission, Opinion of the Data Ethics Commission, January, 22 2020, available at: https://www.bmiv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.html?__blob=publicationFile&cid=7E543D8123E46503CA21DB68AB9A55F7.2_cid334?nn=11678512>, p. 177 [accessed on June 03, 2020].
- 47 The risk assessment procedure and the risk matrix have taken inspiration from: Cf fn. 46; Martini, Mario, Kontrollsystem für algorithmenbasierte Entscheidungsprozesse, 2019, available at: https://www.uni-speyer.de/fileadmin/Lehrstuehle/Martini/2019_Gutachten_GrundlageneinesKontrollsystemendgueltig.pdf>; Zweig, Katharina A., Algorithmische Entscheidungen: Transparenz und Kontrolle, Analysen & Argumente (338),

-
- January 2019, available at:
<https://www.kas.de/documents/252038/4521287/AA338+Algorithmische+Entscheidungen.pdf/533ef913-e567-987d-54c3-1906395cdb81?version=1.0&t=1548228380797> [accessed on June 03, 2020].
- 48 Uber, How Uber's dynamic pricing model works, available at: <https://www.uber.com/en-AE/blog/uber-dynamic-pricing-model/> [accessed on June 08, 2020].
- 49 Cf Chatbot Bobbi, available at: <https://service.berlin.de/chatbot/chatbot-bobbi-606279.php> [accessed on June 08, 2020].
- 50 Cf fn. 9.
- 51 PwC, 2018 AI Predictions, 8 Insights to shape Business Strategy, 2018, available at:
<https://www.pwc.es/es/home/assets/ai-predictions-2018-report.pdf>, p. 18 [accessed on June 03, 2020].
- 52 Whittaker, Meredith/Crawford, Kate/Dobbe, Roel et al., AI Now Report 2018, December 2018, available at:
https://ainowinstitute.org/AI_Now_2018_Report.pdf, p. 22 [accessed on June 03, 2020].
- 53 Buchholtz, Gabriele, Artificial Intelligence and Legal Tech: Challenges to the Rule of Law, in: Wischmeyer, Thomas/Rademacher, Timo (edt) Regulating Artificial Intelligence, Cham: Springer 2019, p. 192.
- 54 Cf fn. 46, p. 74.
- 55 Nassar, Mohamed/Salah, Khaled/ur Rehman Muhammad Habib/Svetinovic, Davoc Blockchain for explainable and trustworthy artificial intelligence, WIREs Data Mining Knowl. Discov., 10(1), October 17, 2019, available at:
<https://doi.org/10.1002/widm.1340>, p. 1 [accessed on June 03, 2020].
- 56 Schmelzer, Ron, Understanding Explainable AI, Forbes, July 23, 2019, available at:
<https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/#120dc44f7c9e> [accessed on June 03, 2020].
- 57 Cf fn. 21, p. 1836.
- 58 Cf fn. 31, pp. 5 et seq.
- 59 Cf fn. 26, p. 4.
- 60 Cf fn. 10, p. 1115.
- 61 Cf Horizon 2020, Research and Innovation programme, available at:
<https://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020> [accessed on June 12, 2020].
- 62 Bussmann, Niklas/Giudici, Paolo/Marinelli, Dimitri/Papenbrock, Jochen, Explainable AI in Credit Risk Management, December 18, 2019, available at: <https://ssrn.com/abstract=3506274>, pp. 3, 15 [accessed on June 03, 2020].
- 63 Ibid p. 10.
- 64 Cf fn. 21, p. 1837.
- 65 Ibid.
- 66 Watson, Grant, Understanding Blockchain Technology, February 14, 2020, available
<https://dev.to/granticusmaximus/understanding-blockchain-technology-56n7> [accessed on June 08, 2020].
- 67 European Parliamentary Research Service (EPRS), Blockchain and the General Data Protection Regulation: Can distributed ledgers be squared with European data protection law?, available at:
[https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634445/EPRS_STU\(2019\)634445_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634445/EPRS_STU(2019)634445_EN.pdf), p. 97 [accessed on June 03, 2020].
- 68 Robbins, Scott, A Misdirected Principle with a Catch: Explicability of AI, Minds and Machines, 2019, available at: <https://doi.org/10.1007/s11023-019-09509-3>, p. 509 [accessed on June 11, 2020].

Bibliography

A

Aircraft Accident Investigation Report, available at:
<http://knkt.dephub.go.id/knkt/ntsc_home/ntsc.htm> [accessed on June 08, 2020].

Aircraft Accident Investigation Bureau Interim Report, available at:
<https://reports.aviation-safety.net/2019/20190310-0_B38M_ET-AVJ_Interim.pdf>
[accessed on June 08, 2020].

Angwin, Julia et al. Machine Bias, May 23, 2016, available at:
<<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>> [accessed on June 07, 2020].

B

Brkan, Maja, Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond, August 1, 2017, available at:
<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3124901>, p. 15 [accessed on June 03, 2020].

Buchholtz, Gabriele, Artificial Intelligence and Legal Tech: Challenges to the Rule of Law, in: Wischmeyer, Thomas/Rademacher, Timo (edt) Regulating Artificial Intelligence, Cham: Springer 2019, p. 192.

Bussmann, Niklas/Giudici, Paolo/Marinelli, Dimitri/Papenbrock, Jochen, Explainable AI in Credit Risk Management, December 18, 2019, available at:
<<https://ssrn.com/abstract=3506274>> [accessed on June 03, 2020].

D

Dastin, Jeffrey, Amazon scraps secret AI recruiting tool that showed bias against women, October 10, 2018, Reuters, available at: <<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>> [accessed on June 08, 2020].

Data Ethics Commission, Opinion of the Data Ethics Commission, January 22, 2020, available at:
<https://www.bmju.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.html?jsessionid=7E543D8123E46503CA21DB68AB9A55F7.2_cid334?n=n=11678512>, p. 177 [accessed on June 03, 2020].

Deeks, Ashley, The Judicial Demand for Explainable Artificial Intelligence, Columbia Law Review 119 (7), 2019, available at: <www.jstor.org/stable/26810851>, p. 1844 [accessed on June 03, 2020].

Delcker, Janosch, In global AI race, Europe pins hopes on ethics, April 25, 2018, available at: <<https://www.politico.eu/article/europe-commission-andrus-ansip-hopes-ethical-approach-will-be-its-edge-in-global-ai-artificial-intelligence-race/>> [accessed on June 12, 2020].

Doshi-Velez, Finale et. al, Accountability of AI Under the Law: The Role of Explanation, available at: <<https://arxiv.org/pdf/1711.01134.pdf>> [accessed on June 03, 2020].

E

Edwards, Lilian/Veale, Michael, 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), 2017, available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2972855>, p. 44 [accessed on June 03, 2020].

European Commission, White Paper: On Artificial Intelligence - A European approach to excellence and trust, February 19, 2020, available at: <https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf>, p. 3 [accessed on June 12, 2020].

European Parliamentary Research Service (EPRS), A governance framework for algorithmic accountability and transparency, April 2019, available at: <[https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU\(2019\)624262_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf)> [accessed on June 03, 2020].

European Parliamentary Research Service (EPRS), Blockchain and the General Data Protection Regulation: Can distributed ledgers be squared with European data protection law?, available at: <[https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634445/EPRS_STU\(2019\)634445_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634445/EPRS_STU(2019)634445_EN.pdf)>, p. 97 [accessed on June 03, 2020].

H

Hacker, Philipp/Krestel, Ralf/Grundmann, Stefan et al., Explainable AI under contract and tort law: legal incentives and technical challenges. Artificial Intelligence Law, 2020, available at: <<https://doi.org/10.1007/s10506-020-09260-6>>, p. 2 [accessed on June 03, 2020].

Horizon 2020, Research and Innovation programme, available at: <<https://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020>> [accessed on June 12, 2020].

L

Langewiesche, William, What really brought down the Boeing 737 Max?, Malfunctions caused two deadly crashes. But an industry that puts unprepared pilots in the cockpit is just as guilty., New York Times Magazine, September 18, 2019, available at:

<<https://www.nytimes.com/2019/09/18/magazine/boeing-737-max-crashes.html>> [accessed on June 03, 2020].

Loria, Kevin, Kurzweil: Human-level AI is coming by 2029, Business Insider, December 29, 2014, available at: <<https://www.businessinsider.com/ray-kurzweil-thinks-well-have-human-level-ai-by-2029-2014-12?r=DE&IR=T>> [accessed on June 08, 2020].

M

Malgieri Gianclaudio/Comand , Giovanni, Why a Right to Legibility of Automated Decision- Making Exists in the General Data Protection Regulation, International Data Privacy Law, 7 (3), 2017, available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3088976>, pp. 243, 245, 250 [accessed on June 03, 2020].

Martini, Mario, Kontrollsystem f r algorithmenbasierte Entscheidungsprozesse, 2019, available at: <https://www.uni-speyer.de/fileadmin/Lehrstuehle/Martini/2019_Gutachten_GrundlageneinesKontrollsystemendgueltig.pdf> [accessed on June 03, 2020].

McKinsey Global Institute, Notes from the AI Frontier Tackling Europe's Gap in Digital and AI, discussion paper, February 2019, available at: <<https://www.mckinsey.com/~media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Tackling%20Europes%20gap%20in%20digital%20and%20AI/MGI-Tackling-Europes-gap-in-digital-and-AI-Feb-2019-vF.ashx>>, pp. 2, 29 [accessed on June 09, 2020].

Mendoza, Isak/Bygrave, Lee A., The Right Not to Be Subject to Automated Decisions Based on Profiling, May 08, 2017, in: Synodinou, Tatiani/Jougleux, Philippe/Markou, Christiana/Prastitou, Thalia (eds.), EU Internet Law: Regulation and Enforcement, Springer, 2017, University of Oslo Faculty of Law Research Paper No. 2017-20, available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2964855> [accessed on June 03, 2020].

N

Nassar, Mohamed/Salah, Khaled/ur Rehman Muhammad Habib/Svetinovic, Davoc Blockchain for explainable and trustworthy artificial intelligence, WIREs Data Mining Knowl. Discov., 10(1), October 17, 2019, available at: <<https://doi.org/10.1002/widm.1340>> p. 1 [accessed on June 03, 2020].

P

PayPal, Information about the PayPal identity check and data exchange with credit agencies (credit check), December 18, 2018, available at: <https://www.paypal.com/de/webapps/mpp/ua/creditchk?locale.x=en_DE> [accessed on June 12, 2020].

PwC, 2018 AI Predictions, 8 Insights to shape Business Strategy, 2018, available at: <https://www.pwc.es/es/home/assets/ai-predictions-2018-report.pdf>, p. 18 [accessed on June 03, 2020].

R

Renda, Andrea, Artificial Intelligence, Ethics, governance, and policy changes, Report of a CEPS Task Force, Centre for European Policy Studies (CEPS) Brussels, February 2019, available at: https://www.ceps.eu/wp-content/uploads/2019/02/AI_TFR.pdf, p. 45 [accessed on June 03, 2020].

Robbins, Scott, A Misdirected Principle with a Catch: Explicability of AI, Minds and Machines, 2019, available at: <https://doi.org/10.1007/s11023-019-09509-3>, p. 509 [accessed on June 11, 2020].

Ross, Casey/Swetlitz, Ike, IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show, STAT+, July 25, 2018, available at: <https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf> [accessed on June 03, 2020].

S

Selbst, Andrew D./Barocas, Solon, The Intuitive Appeal of Explainable Machines, Fordham Law Review 87 (3), December 2018, available at: <http://dx.doi.org/10.2139/ssrn.3126971>, p. 1085 [accessed on June 03, 2020].

Selbst, Andrew D./Powles, Julia, Meaningful information and the right to explanation, International Data Privacy Law 7 (4), November 2017, available at: <https://academic.oup.com/idpl/article/7/4/233/4762325> [accessed on June 03, 2020].

Schmelzer, Ron, Understanding Explainable AI, Forbes, July 23, 2019, available at: <https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/#120dc44f7c9e> [accessed on June 03, 2020].

T

Tobey, Danny, Explainability: Where AI and liability meet, February 25, 2019, available at: <https://www.dlapiper.com/de/germany/insights/publications/2019/02/explainability-where-ai-and-liability-meet/> [accessed on June 03, 2020].

U

Uber, How Uber's dynamic pricing model works, available at: <https://www.uber.com/en-AE/blog/uber-dynamic-pricing-model/> [accessed on June 08, 2020].

U.S. Department of the Treasury, Opportunities and Challenges in Online Marketplace Lending, May 10, 2016, available at: https://www.treasury.gov/connect/blog/Documents/Opportunities_and_Challenges_in_Online_Marketplace_Lending_white_paper.pdf>, p. 20 [accessed on June 08].

Y

Yeung, Karen, Responsibility and AI, Council of Europe Study, September 2019, available at: <https://rm.coe.int/responsability-and-ai-en/168097d9c5>>, p. 21 [accessed on June 03, 2020].

W

Wachter, Sandra et al., Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, Harvard Journal of Law and Technology, 31 (2), 2018, available at: <https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf>>, p. 841 [accessed on June 03, 2020].

Watson, Grant, Understanding Blockchain Technology, February 14, 2020, available <https://dev.to/granticusmaximus/understanding-blockchain-technology-56n7>> [accessed on June 08, 2020].

Whittaker, Meredith/Crawford, Kate/Dobbe, Roel et al., AI Now Report 2018, December 2018, available at: https://ainowinstitute.org/AI_Now_2018_Report.pdf>, p. 22 [accessed on June 03, 2020].

Z

Zweig, Katharina A., Algorithmische Entscheidungen: Transparenz und Kontrolle, Analysen & Argumente (338), January 2019, available at: <https://www.kas.de/documents/252038/4521287/AA338+Algorithmische+Entscheidungen.pdf/533ef913-e567-987d-54c3-1906395cdb81?version=1.0&t=1548228380797>> [accessed on June 03, 2020].

Acknowledgements

V29 Legal would like to express a special gratitude to Viola Zollitsch, LLM, for her invaluable support in putting this Contribution Paper together and dedication to the topic. In addition, V29 Legal is grateful to Dr Jochan Papenbrock, CEO of the Frankfurt-based Fintech Firamis and consortium partner of the EU Horizon2020 project FIN-TECH, for providing input with regards to explainable AI and Firamis-technology.





Prof. Dr. Christian Duve

Partner

Olga Hamama

Partner



V29 Legal – Duve Hamama Rechtsanwälte PartG mbB

TechQuartier
Platz der Einheit 2
60327 Frankfurt am Main
Germany



+49 (0)69 26 40 17 59



office@v29-legal.com



www.v29-legal.com