Artificial Intelligence and Law
https://doi.org/10.1007/s10506-020-09260-6

ORIGINAL RESEARCH

Explainable AI under contract and tort law: legal incentives and technical challenges

Philipp Hacker[1] · Ralf Krestel[2] · Stefan Grundmann[1] · Felix Naumann[2]

Abstract
This paper shows that the law, in subtle ways, may set hitherto unrecognized incentives for the adoption of explainable machine learning applications. In doing so, we make two novel contributions. First, on the legal side, we show that to avoid liability, professional actors, such as doctors and managers, may soon be legally compelled to use explainable ML models. We argue that the importance of explainability reaches far beyond data protection law, and crucially influences questions of contractual and tort liability for the use of ML models. To this effect, we conduct two legal case studies, in medical and corporate merger applications of ML. As a second contribution, we discuss the (legally required) trade-off between accuracy and explainability and demonstrate the effect in a technical case study in the context of spam classification.

Keywords Explainability · Explainable AI · Interpretable machine learning · Contract law · Tort law · Explainability-accuracy trade-off · Medical malpractice · Corporate takeovers

✉ Philipp Hacker
philipp.hacker@rewi.hu-berlin.de

Ralf Krestel
ralf.krestel@hpi.uni-potsdam.de

Stefan Grundmann
stefan.grundmann@rewi.hu-berlin.de

Felix Naumann
felix.naumann@hpi.uni-potsdam.de

[1] Law Department, Humboldt University of Berlin, Unter den Linden 6, 10099 Berlin, Germany

[2] Hasso Plattner Institute, University of Potsdam, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany

Published online: 19 January 2020

1 AI and the law

Machine learning is the most prominent and economically relevant instantiation of artificial intelligence techniques (Royal Society 2017). While the potential of machine learning for economic decision-making contexts, from lending decisions to employment candidate selection, has long been recognized and implemented in the digital economy (Witten et al. 2016; Cowgill 2017), algorithmic models are increasingly emerging as the subject of an intense policy debate surrounding the legality of, and liability for, advanced machine learning applications (Reed et al. 2016; Calo 2016; Selbst 2019).

In this debate, the problem of explainability of decisions reached with the help of machine learning tools has assumed a central role at the intersection of artificial intelligence and the law. Today, most of this debate is focused, on the legal side, on data protection law (Wachter et al. 2017; Selbst and Powles 2017; Doshi-Velez 2017; Goodman and Flaxman 2016; Malgieri and Comandé 2017; Wischmeyer 2018). An intense scholarly debate has erupted over whether the General Data Protection Regulation (GDPR) includes a right to an explanation of automated decisions. In this paper, we seek to broaden the scope of this debate along two dimensions. First, we aim to show that explainability is an important legal category not only in data protection law, but also in contract and tort law. With legal requirements under the GDPR being largely unclear at the moment, it seems fruitful to inquire into the role of explainability in other legal areas. In fact, as we argue, contract and tort law, not data protection law, may eventually impose legal requirements to use explainable machine learning models. An upshot of this discussion is that the trade-off between explainability and accuracy, much discussed in the technical literature (Mori and Uchihira 2018; Rudin 2019), sits at the heart of contract and tort liability for the use of AI systems.

Second, we empirically investigate that trade-off in an experiment on spam classification to show that the precise calibration of the trade-off crucially turns on the choice of the machine learning model, and the task it is employed for. This, in turn, has repercussions for the legal analysis concerning liability.

This paper therefore aims to show that it would be a mistake to assume a facile dichotomy of technology as an enabling and the law as a limiting factor for the deployment of machine learning technology in economic sectors. Rather, as we aim to highlight, artificial intelligence and the law are interwoven and, often, mutually reinforcing: while it is true that the law does set certain *limits* on the use of artificial intelligence, it also, and often in the same context, may in fact *require* the use of explainable machine learned models for economic agents to fulfill their duties of care. For example, to avoid liability for medical malpractice, doctors may soon have to rely on machine learning technology if a model can diagnose a disease (statistically) better than any human eye or brain could; similarly, companies may be obliged to incorporate machine learned models into their complex risk assessment strategies, if they are superior to traditional tools, in order to fulfill regulatory requirements for effective risk management. Machine learning research, in turn, can benefit from such legal requirements not only by

increasing exposure of models to different situational contexts (providing additional opportunities for learning), but also by receiving incentives for the development of novel types of algorithms, such as explainable AI and user-friendly interfaces. Allowing this type of engagement between the user and the model will be crucial for human users, and society at large, to develop the trust in AI necessary for large-scale application and adoption (Lipton 2018; Ribeiro et al. 2016).

The remainder of this article is structured as follows: In Sect. 2, we introduce the discussion surrounding explainability in AI and the law. Section 3 is dedicated to two distinct case studies from the perspective of law on medical and corporate applications of AI. Section 4 takes the perspective of computer science discussing the trade-off between explainability and accuracy. Finally, Sect. 5 concludes the article.

2 Explainability in AI and the law

Developers of machine learning applications routinely have to make a choice between different types of models to optimize a task. Such types range from simple decision trees all the way to deep neural networks. To meet a regulatory burden, or to avert liability, decision makers have to choose a model that works well in the context at hand. To this extent, the law is agnostic as to the underlying model.

However, in recent years, explainability has emerged as a key factor informing model choice both from a technical and from a legal point of view. In technical terms, as we show in greater detail in Sect. 4, different models vary with respect to their degree of explainability. While there is no uniformly accepted definition of explainability in machine learning, models can roughly be grouped into those that are interpretable ex ante and those that can be explained only ex post (Lipton 2018; Rudin 2019). Linear regression models, for example, are typically interpretable ex ante as the regression coefficients provide an understanding of the respective weights of the features used to make a prediction (Lapuschkin et al. 2019). Deep neural networks, on the other hand, are typically so complex that specific weights cannot be determined for individual features in a global manner, i.e., for all possible predictions (Lapuschkin et al. 2019; Lipton 2018; Rudin 2019). However, a number of studies have shown that it is often possible to identify ex post the features (input variables) that are responsible for a specific output (Samek et al. 2017; Montavon et al. 2017; Bach et al. 2015; Lapuschkin et al. 2019; Ribeiro et al. 2016).

In legal terms, explainability has almost uniquely been discussed as a possible requirement under data protection law (see refs in Sect. 1). Particularly concerning the GDPR, scholars are divided over whether the regulation contains a right to explanation. The debate has focused on two provisions. Art. 22(3) GDPR holds that, in certain cases of automated processing, "the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision". Recital 71 of the GDPR, in discussing Art. 22(3) GDPR, famously states that the safeguards should, inter alia, include "the right [...] to obtain an explanation of the decision reached after such assessment". Since the right to explanation

is contained only in the (non-binding) recital, and not in the binding text of Art. 22(3) GDPR, most scholars agree that a right to an explanation of individual decisions, which could include global or local explanations, does not follow from Art. 22(3) GDPR (Wachter et al. 2017; Goodman and Flaxman 2017; Selbst and Powles 2017; Wischmeyer 2018).

This is not the end of the story for EU data protection law, however. Art. 15(1)(h) GDPR stipulates that, in the case of automated processing in the sense of Art. 22(1) GDPR, the controller must provide "meaningful information about the logic involved". Here, scholars are deeply divided over the implications of this provision. Some argue that it only refers to the general structure and architecture of the processing model, but that explanations about individual decisions or concrete weights and features of the model need not be provided (Wachter et al. 2017; Malgieri and Comandé 2017; Party 2017). Others suggest that information can only be meaningful if it helps the data subject to exercise her rights under Art. 22(1) and (3) GDPR, which includes the right to contestation of the decision. Therefore, if specific explanations of the decision, including the weights and factors used to reach it, are necessary to check the accuracy of the protection and to potentially challenge its correctness, such information needs to be provided (Selbst and Powles 2017). While this interpretation indeed affords the advantage of integrating Art. 15(1)(h) into the purpose of Art. 22 GDPR, to which it explicitly refers, scholarly consensus is far from emerging and the question will likely have to be settled by the Court of Justice of the European Union. Furthermore, the entire provision hinges on the existence of automated processing in the sense of Art. 22(1) GDPR. However, for this to apply, the decision must be "based solely on automated processing". Hence, on the face of it, any kind of human intervention in the decision exempts the controller from the constraints of Art. 22 and Art. 15(1)(h) GDPR (Party 2017; Selbst and Powles 2017; Wachter et al. 2017). Therefore, explainability faces, at best, a highly uncertain future under the GDPR.

This is precisely why this article turns to contract and tort law to discuss the relationship between explainability and liability in these areas. In these domains, there are two ways in which the law does affect the adoption of machine learning models in practice. First, the question arises if certain actors indeed *have to* use ML models, instead of or at least complementing human decision making, to *avoid liability* if the model consistently outperforms humans (see below, under Sects. 3.1 and 3.2). We shall investigate this matter in two case studies: liability of medical doctors for failing to use ML diagnostics; and liability for corporate managers for failing to use ML company valuation tools in Mergers and Acquisitions (M&A). As the case studies show, the degree to which the law compels the adoption of ML tools is, inter alia, a function of their explainability. Second, liability for the selection of the precise machine learning model crucially hinges on the relationship between explainability and accuracy (or other performance metrics). However, as we show in Sect. 4, in state-of-the-art models explainability and performance in terms of predication accuracy are often proportionally inverse: easily explainable models may perform poorly, while the currently best performing models are so highly complex that even expert developers can explain neither intuitively nor technically how the good results come about. Since this trade-off is also at the heart of questions of contract and

tort liability, the technical explainability-accuracy trade-off replicates in the legal domain. We now turn to the case studies to elaborate these points further.

3 Case studies

Our two case studies, on medical diagnostics/malpractice and on corporate valuation/the business judgment rule, analyze recent advances in ML prediction tools from a legal point of view. Importantly, they go beyond the current discussion around the data protection requirements of explainability to show that explainability is a crucial, but overlooked, category for the assessment of contractual and tort liability concerning the use of AI tools. We follow up in Sect. 4 on explainability proper, additionally implementing an exemplary spam classification, to discuss the trade-off between accuracy and explainability from both a technical and a legal perspective.

3.1 Medical diagnostics

Medical diagnostics is a field where technology and the law patently interconnect. The medical standard of care defines contractual and tort liability for medical malpractice. However, this standard is itself shaped by state-of-the-art technology. If doctors fail to use novel, ML-driven methods, which are required by the applicable standard of care, liability potentially looms large. Conversely, if they apply models that result in erroneous predictions, they are equally threatened by liability. Importantly, both questions are intimately connected to explainability, as we argue below.

3.1.1 The rise of ML diagnostics

Indeed, ML models using a variety of techniques are increasingly entering the field of medical diagnostics and treatment [overview in Topol (2019)]. For example, researchers developed a model, based on reinforcement learning, for the diagnosis and optimal treatment of patients with sepsis in intensive care. The model, termed "AI clinician", analyzes the patient's state and selects appropriate medication doses (Komorowski et al. 2018). On average, the AI clinician outperformed human intensive care experts. In a large validation data set, different from the training data set, patients' survival rate was highest when the actual doses administered by human clinicians matched the AI clinician's predictions.In another recent study, a deep neural network was trained to detect Alzheimer disease based on brain scans (Ding 2018). Not only did the network outperform human analysts by an important margin; it also detected the disease an average 75 months before the final clinical diagnosis. This makes particularly effective early treatment medication available to patients much earlier. Supra-human performance could also be established in the detection of retinal diseases (De Fauw 2018) and in heart attack prediction (Weng et al. 2017).

ML is increasingly used in cancer diagnosis and treatment, too (Kourou et al. 2015). This is often due to advances in pattern recognition in images, a field in

which deep neural networks perform particularly well. Researchers trained a convolutional neural network (CNN) to identify skin cancer on the basis of images of skin lesions (Esteva et al. 2017). The CNN matched the performance of 21 board-certified dermatologists in the classification of two varieties of skin cancer (the most common and the deadliest version of skin cancer). Similarly, a Chinese ML was reported to have beaten a team of 15 experienced doctors in brain tumor recognition (Press 2018). However, not all that glitters is gold: IBM's Watson had, according to news reports, difficulties in correctly identifying cancer patients (Ross and Swetlitz 2018), and hospitals were reported to cut back collaborations with IBM (Jaklevic 2017). This shows that rigorous field validation is necessary before ML models can be safely used in medical contexts—an issue that informs our analysis of the legal prerequisites for the adoption of such models.

3.1.2 Legal liability

As the previous section has shown, predictions used by medical AI models are not, of course, fully accurate in every case. Hence, we shall ask what factors determine whether such a potentially erroneous model may be used by a medical doctor without incurring liability. Furthermore, with ML technology approaching, and in some cases even surpassing, human capacities in medical diagnostics, the question arises whether the failure to use such models may constitute medical malpractice. The relevant legal provisions under contract and tort law differ from country to country. We offer a legal analysis in which we generally refer to German and US law in particular; nevertheless, general normative guidelines can be formulated.

Adoption For the sake of simplicity, we assume that all formal requirements for the use of the ML model in medical contexts are met. In April 2018, for example, the US Food and Drug Administration (FDA) approved IDxDR, an ML tool for the detection of diabetes-related eye disorders (US Food and Drug Administration 2018). While liability for the adoption of new medical technology is an obvious concern for medical malpractice law (Katzenmeier 2006; Greenberg 2009), the issue has, to our knowledge, not been discussed with an explicit focus on explainability of ML models. The related but different question whether the avoidance of legal liability *compels* the adoption of such a model has rarely been discussed in literature (Froomkin 2018; Greenberg 2009) and, to our knowledge, not at all by the courts. The answer to both questions crucially depends on whether it is considered negligent, under contractual and tort liability, (not) to use ML models during the treatment process. This, in turn, is determined by the appropriate medical standard of care.

Generally speaking, healthcare providers, such as hospitals or doctor's practices, cannot be required to always purchase and use the very best products on the market. For example, when a new, more precise version of an x-ray machine becomes available, it would be ruinous for healthcare providers to be compelled to always immediately buy such new equipment. Therefore, they must be allowed to rely on their existing methods and products as long as these practices guarantee a satisfactory level of diagnostic accuracy, i.e., as long as they fall within the "state of the art" (Hart 2000). However, as new and empirically better products become available,

the minimum threshold of acceptable accuracy moves upward. Otherwise, medical progress could not enter negligence norms. Hence, the content of the medical standard, whose fulfilment excludes negligence, is informed not only by experience and professional acceptance, but also (and in an increasingly dominant way) by empirical evidence (Hart 2000; Froomkin 2018).

Importantly, therefore, the acceptable level of accuracy could change with the introduction of new, more precise, ML driven models. Three criteria, we argue, should be met for this to be the case. First, the use of the model must not, in itself, lead to medical malpractice liability. This criterion, therefore, addresses our first question concerning liability for the positive use of ML technology in medicine. To rely on the model, there must be a significant difference between the performance of the model and human-only decision making in its absence. This difference must be shown consistently in a number of independent studies and be validated in real-world clinical settings—which is often lacking at the moment (Topol 2019). The superiority of the model cannot be measured only in terms of its accuracy (i.e., the ratio of correct over all predictions); rather, other performance metrics, such as sensitivity (a measure of false negatives)[1] or specificity (a measure of false positives),[2] also need to be considered. Depending on the specific area, a low false positive or false negative rate may be equally desirable, or even more important, than superior accuracy (Froomkin 2018; Topol 2019; Caruana 2015). For example, false negatives in tumor detection will mean that the cancer can grow untreated—quite likely the worst medial outcome. The deep learning model for detecting Alzheimer disease, for example, did have both higher specificity and sensitivity (and hence higher accuracy) than human radiologists (Ding 2018). The superiority of the novel method must, in addition, be plausible for the concrete case at hand to be legitimate in that instance. Under German law, for example, new medical methods meet the standard of care if the marginal advantages vis-à-vis conventional methods outweigh the disadvantages for an individual patient (Katzenmeier 2006); the same holds true for US law (Greenberg 2009).

This has important implications for the choice between an explainable and a non-explainable model: the non-explainable model may be implemented only if the marginal benefits of its use (improved accuracy) outweigh the marginal costs. This depends on whether the lack of explainability entails significant risks for patients, such as risks of undetected false negative treatment decisions. As Ribeiro et al. (2016) rightly argue, explainability is crucial for medical professionals to assess whether a prediction is made based on plausible factors or not. The use of an explainable model facilitates the detection of false positives and false negative classifications, because it provides medical doctors with reasons for the predictions, which can be critically discussed (Lapuschkin et al. 2019; Lipton 2018) (see, in more detail, the next section). However, this does not imply

[1] Sensitivity is defined as the ratio of true positives over all positives, i.e., over the sum of true positives and false negatives. Sensitivity is also called recall in ML contexts.

[2] Specificity is defined as the ratio of true negatives over all negatives, i.e., over the sum of true negatives and false positives.

that explainable models should always be chosen over non-explainable models. Clearly, if an explainable model performs equally well as a non-explainable one, the former must be chosen [for examples, see Rudin (2019) and Rudin and Ustun (2018)]. But, if explainability reduces accuracy—which need not necessarily be the case, see Rudin (2019) and below, Sect. 4—the choice of an explainable model will lead to some inaccurate decisions that would have been accurately taken under a non-explainable model with superior accuracy. Therefore, in these cases, doctors must diligently weigh the respective marginal costs and benefits of the models. In some situations, it may be possible, given general medical knowledge, to detect false predictions even without having access to the factors the model uses. In this case, the standard of care simply dictates that the model with significantly superior accuracy should be chosen. If, however, the detection of false predictions, particularly of false negatives, requires or is significantly facilitated by an explanation of the algorithmic model, the standard of care will necessitate the choice of the explainable model. Arguably, this will often be the case: it seems difficult to evaluate the model predictions in the field without access to the underlying factors used, precisely because it will often be impossible to say whether a divergence from traditional medical wisdom is due to a failure of the model or to its superior diagnostic qualities.

Hence, general contract and tort law significantly constrains the use of non-explainable ML models—arguably, in more important ways than data protection law. Only if the balancing condition (between the respective costs and benefits of accuracy and explainability) is met, the use of the model should be deemed generally legitimate (but not yet obligatory).

Second, for the standard of care to be adjusted upward, and hence the use of a model to become obligatory, it must be possible to integrate the ML model smoothly into the medical workflow. High accuracy does not translate directly into clinical utility (Topol 2019). Hence, a clinical suitability criterion is necessary since ML models pose particular challenges in terms of interpretation and integration into medical routines, as the Watson case showed. Again, such smooth functioning in the field generally includes the explainability of the model to an extent that decision makers can adopt a critical stance toward the model's recommendations (see, in detail, below, Sect. 3.1.2, Use of the model). Note that this criterion is independent of the one just discussed: it does not involve a trade-off with accuracy. Rather, explainability per se is general pre-condition for the duty (but not for the legitimacy) to use ML models: while it may be legitimate to use a black box model (our first criterion), there is, as a general principle, no duty to use it. A critical, reasoned stance toward a black box model's advice is difficult to achieve, and the model will be difficult to implement into the medical workflow. As a general rule, therefore, explainability is a necessary condition for a duty to use the model, but not for the legitimacy of the use of a model. The clinical suitability criterion is particularly important in medical contexts where the consequences of false positive or false negative outcomes may be particularly undesirable (Caruana 2015; Zech et al. 2018; Rudin and Ustun 2018). Therefore, doctors must be in a position to check the reasons for a specific outcome. Novel techniques of local explainability of even highly complex models may provide for such features (Ribeiro et al. 2016; Lapuschkin et al. 2019).

Exceptionally, however, the use of black box models with supra-human accuracy may one day become obligatory if their field performance on some dimension (e.g., sensitivity) is exceptionally high (e.g.,> 0.95) and hence there is a reduced need for arguing with the model within its high performance space (e.g., avoidance of false negatives). While such extremely powerful, non-explainable models still seem quite a long way off in the field (Topol 2019), there may one day be a duty, restricted to their high performance domain, to use them if suitable routines for cases of disagreement with the model can be established. For example, if a doctor disagrees with a close-to-perfect black box model, the case may be internally reviewed by a larger panel of (human) specialists. Again, if these routines can be integrated into the medical workflow, the second criterion is fulfilled.

Finally, third, the cost of the model must be justified with respect to the total revenue of the healthcare provider for the latter to be obliged to adopt it (Froomkin 2018). Theoretically, licensing costs could be prohibitive for smaller practices. In this case, they will, however, have to refer the patient to a practice equipped with the state-of-the-art ML tool.

While these criteria partly rely on empirical questions (particularly the first one), courts are in a position to exercise independent judgment with respect to their normative aspects (Greenberg 2009). Even clinical practice guidelines indicate, but do not conclusively decide, a (lack of) negligence in specific cases (Laufs 1990).

In sum, to avoid negligence, medical doctors need not resort to the most accurate product, including ML models, but to state-of-the-art products that reach an acceptable level of accuracy. However, this level of accuracy should be adjusted upwards if ML models are shown to be consistently superior to human decision making, if they can be reasonably integrated into the medical workflow, and if they are cost-justified for the individual health-care provider. The choice of the concrete model, in turn, depends on the trade-off between explainability and accuracy, which varies between different models.

Use of the model Importantly, even when the use of some ML model is justified or even obligatory, there must be room for reasoned disagreement with the model. Concrete guidelines for the legal consequences of the use of the model are basically lacking in the literature. However, we may draw on scholarship regarding evidence-based medicine to tackle this problem. This strand of medicine uses statistical methods (for example randomized controlled trials) to develop appropriate treatment methods and displace routines based on tradition and intuition where these are not upheld by empirical evidence (Timmermans and Mauck 2005; Rosoff 2001). The use of ML models pursues a similar aim. While it does, at this stage, typically not include randomized controlled trials (Topol 2019), it is also based on empirical data and seeks to improve on intuitive treatment methods.

Of course, even models superior to human judgment on average will generate some false negative and false positive recommendations. Hence, the use of the model should always only be part of a more comprehensive assessment, which includes and draws on medical experience (Froomkin 2018). Doctors, or other professional agents, must not be reduced to mere executors of ML judgments. If there is sufficient, professionally grounded reason to believe the model is wrong in a particular case, its decision must be overridden. In this case, such

a departure from the model must not trigger liability—irrespective of whether the model was in fact wrong or right in retrospect. This is because negligence law does not sanction damaging outcomes, as strict liability does, but attaches liability only to actions failing the standard of care. Hence, even if the doctor's more comprehensive assessment is eventually wrong and the model prediction was right, the doctor is shielded from medical malpractice claims as long as his reasons for departing from the model were based on grounds justified on the basis of professional knowledge and behavior. Conversely, *not* departing a wrong model prediction would breach the standard of care if, and only if, the reasons for departure were sufficiently obvious to a professional (Droste 2018). An example may be an outlier case, which, most likely, did not form part of the training data of the ML model, cf. Rudin (2019). However, as long as such convincing reasons for model correction cannot be advanced, the model's advice may be heeded without incurring liability, even if it was wrong in retrospect (Thomas 2017). This is a key inside of the scholarship on evidence-based medicine (Wagner 2018). The reason for this rule is that, if the model indeed is provably superior to human professional judgment, following the model will on average produce less harm than a departure from the model (Wagner 2018). Potentially, a patient could, in these cases, direct a product liability claim against the provider of the ML model (Droste 2018).

Particularly in ML contexts, human oversight, and the possibility to disagree with the model based on medical reasons, seems of utmost importance: ML often makes mistakes humans would not make (and vice versa). Hence, the possibility of reasoned departure from even a supra-human model creates a machine-human-team, which likely works better than either machine or human alone (Thomas 2017; Froomkin 2018). Importantly, the obligation to override the model in case of professional reasons ensures that blindly following the model, and withholding individual judgment, is not a liability-minimizing strategy for doctors.

3.1.3 Short summary of case study 1

Summing up, ML models are approaching, and in some cases even surpassing, human decision-making capacity in the medical realm. However, the legal standard of care should be adjusted only upward by the introduction of such models if (1) they are proven to be consistently outperforming professional actors and other models, (2) they can be smoothly integrated into the medical workflow, and (3) its use is cost-justified for the concrete healthcare provider. Conditions (1) and (2) are intimately related to the explainability of the model. If Condition (1) is justified, the use of the model per se does not trigger medial liability. If all three are justified, failure to use it leads to liability. Finally, when a model's use is justified and even when it is obligatory, doctors are compelled, under negligence law, to exercise independent judgment and may disagree with the model, based on professional reasons, without risking legal liability.

3.2 Mergers and acquisition

Similar issues arise when machine learning models are used to predict the transaction value of companies (Zuo et al. 2017). Again, the standard of care managers have to employ with respect to due diligence and other preparatory steps to a merger, but as well in the decision proper, are themselves shaped by state-of-the-art technology, namely information retrieval techniques. If managers fail to use novel, ML-driven methods, which are required by the applicable standard of care, the question of liability arises. One particular feature in transactional contexts (M&A) has to be seen, however, in the high transaction costs of mergers and acquisitions. Therefore, the business judgment rule to be applied in this context is seen to require rather conservative models. This might imply that methods with low false positive rates (recommending a transaction) should be preferred for matters of risk analysis and due diligence. Once ML models surpass a certain threshold of predictive value, reducing false positive rates, it may become mandatory for directors to use these models to avail themselves of the business judgment rule, i.e., the rule that protects managers from liability if their predictions turn out to be wrong. It may even be that also false negative rates—even though much less likely to raise concerns—could become a concern for managers, albeit more from a managerial success, not a liability point of view.

3.2.1 The rise of ML valuation tools

Indeed, ML models using a variety of techniques are increasingly entering also the field of mergers and acquisitions (Jiang 2018). A first set of papers analyses due to which factors mergers have come about in the past, using machine learning. For example, a group of researchers developed a model, based on latest machine learning applied to so-called earnings call transcripts, analyzed the role of different types of corporate culture on the number and the rapidity of mergers (Li 2018)—however, not yet reporting on later failure of the mergers analyzed. Earnings call transcripts are summaries of conversations between CEOs, in part also CFOs with analysts that (indirectly) show what is seen as being essential at the most professional conversation level. The results found incrementally suggested that corporate culture focusing on innovation can be distinguished from corporate culture focusing on quality and that two conclusions can be drawn. The first conclusion is that firms with a corporate culture focusing on innovation are more likely to be acquirers than firms with a corporate culture focusing on quality. In the second conclusion, authors find that firms with the same or a more similar corporate culture (from the same of both categories named or also within them) tend to be merged more often and with less transaction costs. While the first conclusion appears to be more descriptive and of less immediate impact on a legal assessment, and while the second set of conclusions seem not really astonishing, at least the core distinction is more refined than normally found in corporate literature. Moreover, transaction costs are certainly one of the key information parameters to be analyzed and taken into account in an assessment under the business judgment rule (see below).

In a second set of articles, researchers focused on particular countries, because of the strong varieties of capitalism element in corporate mergers, for instance Japan, partly still before machine learning, partly with machine learning based on a large-scale data set concentrating on the Japanese corporate setting (Shibayama 2008; Shao et al. 2018).

Even the earlier study would potentially largely profit from machine learning. The studies are different, but both focus on rather substantial sets of data, one on cash flow analyses—still in some part to be conducted. The other study came to the (non-machine learning based) conclusion that knowledge transfer—so important as a motive for mergers—leads indeed to increased innovation (with respect to the development of drugs). According to authors, however, this is not universally true, indeed not so much in the case of a merger of equals (at least in Japan), but of non-equals (big and small), and moreover innovation then takes place mainly in the first few years (two to three). The data set used in the study might form an ideal test case for a comparative assessment of the respective strengths and reliabilities of conventional analyses in the area of mergers (focusing on one industry in one big country) and machine-based analyses.

However, in practical, but also in legal terms, the development of machine learning in mergers and acquisitions seems to lag behind that in medical care (and malpractice) rather considerably. This is implied by the relatively low number of empirical publications and applications in this field. Hence, it is safe to infer that by sheer numbers, mergers supported by machine learning account only for a very small fraction of mergers. A duty to use these techniques is therefore imaginable only in the future. Moreover, the techniques used so far use both supervised and unsupervised techniques of learning, with only very recent discussion and evaluation of the comparative advantages (Jiang 2018). Finally, the decisive question of how high the failure rate is when using results with and without machine learning is still basically undecided. This shows that rigorous field validation is necessary before ML models can be safely used in corporate/merger contexts—an issue that informs our analysis of the legal prerequisites for the adoption of such models. This also shows that designs of machine learning still have to be better adapted (also) to those questions that are legally decisive, namely to the question of failure that is core for the most important legal issue, i.e., liability (of course, also a core question in economic terms).

3.2.2 Legal liability: the business judgment rule

With ML technology approaching, and potentially (in the future) even surpassing, human capacities in due diligence and other decisional steps to be taken in mergers and acquisitions, namely in the choice and evaluation of a target firm, the question arises whether the failure to use such models may constitute a violation of the duty of care. For setting the standard of care in such decisions, its attenuation via the so-called business judgment rule has to be taken into account. The relevant legal provisions under corporate law (again) differ from country to country. We again offer a legal analysis in which we generally refer to German and US law in particular; nevertheless, general normative guidelines can be formulated.

Adoption For setting the legally required parameters correctly, a model has to start from the rather nuanced legal basis of such decision taking—in which avoiding false positives would seem to be more required than avoiding false negatives. Contrary to what has been explained for ML in medicine, the use of ML does not require permission by the competent authority, but is within the decision power of managers, while the situation is similar for the second point. Again, the question whether the avoidance of legal liability compels the adoption of such a model has (to our knowledge) not been discussed in literature nor in the courts. The answer to this question crucially depends on whether it is considered negligent under corporate law standards, such as Section 93 of the German Stock Corporation Code (Aktiengesetz).

Germany and the US both distinguish between a very demanding standard of professional care boards and managers have to exercise—Germany in an even more outspoken way (Hopt and Roth 2015)—, and an application of the so-called business judgment rule. Indeed, both jurisdictions also attenuate the strict standard of care in genuine business decisions via application of the business judgment rule. This rule has been introduced also in Germany—on the model of US law—in section 93 para. 1(2) of the Stock Corporation Code, different, for instance, from UK law (Kraakman 2017; Varzaly 2012). In principle, this rule is aimed at giving managers more lee-way in such decisions and foster courageous decision taking in genuine business matters, giving rather extended discretion to boards/managers whenever the decision made does not appear to be utterly mistaken. This also guards against hindsight bias in evaluating managerial decisions. The rule nevertheless has its limits. Among the most important are the prerequisite of legality—legal limits have to be respected, namely those of penal law as in the famous Ackermann case of Deutsche Bank (Kraakman 2017)—, but additionally—and still more important in our context—the informational basis for a decision has to be retrieved in a sufficiently professional way. Thus courts restrict scrutiny in content to the outer limits, but substitute such scrutiny by stricter procedural rules—namely on information retrieval. A core limiting factor for the adoption and use of ML is the regulation of the structure of a merger (in the narrow sense) or acquisition (takeover) under European, German and US Law. It requires not only full information of the board (within the business judgment rule), but also ample disclosure and co-decision taking by other bodies within the companies—thus making explainability of the process and the results reached via ML paramount (Grundmann 2012).

Given this legal basis, one can, in principle, refer to the tree criteria developed above for medical malpractice (and their discussion) also for the corporate merger setting. In the corporate setting, it would, first, also not constitute improper use of resources if management recurred to ML techniques only once a significant difference between the performance of the model and human-only decision making in its absence can be shown—again in a number of independent studies, also validated in real world settings. In contrast to the medical context, however, false positives and false negatives tend in principle not to have the same weight in corporate mergers and acquisitions. Rather, for the reasons given (transaction costs), a low false positive rate would in principle have to be the prime standard and target of such assessment. Again, this shows that explainability is key: to evaluate whether a prediction

might be incorrectly positive, an explanation will generally be required. Hence, the trade-off between accuracy and explainability explored in the section on medical malpractice reproduces in the corporate setting, too. Second, for the standard of care to be adjusted upward, it must be possible to integrate the ML model smoothly into the procedure of a merger and acquisition undertaking, namely disclosure and possibility of independent decision taking in several distinct bodies. Again, smooth functioning therefore includes the explainability of the model to an extent that decision makers can adopt a critical stance toward the model's recommendations. Third, the cost of ML is, however, less important. Companies need not refrain from a transaction if they cannot reasonably afford the costs of ML—as no third party is affected by the decision to an intolerable extent. Rather, management has to make an overall cost-benefit analysis, in which the cost of ML will be one, but only one, factor. Given the high stakes usually at play in corporate mergers and acquisitions, the cost of ML will rarely be dispositive.

Use of the Model In the corporate setting, questions of use of the model are shaped still more prominently by the factual basis of the subject matter and by the structure of legal rules, namely the business judgment rule. These peculiarities render the results reached above for medical malpractice still more important and unquestionable in corporate law. First, mergers and acquisitions are generally not very successful in the long-term (Jiang 2018)—the failure rate after five years from when the transaction is carried through amounts to up to 50% according to some studies. Moreover, the business judgment rule allows not only discretion whenever proper research (information retrieval) has been accomplished, but it also imposes an evaluation of the information found by the managers. Thus, in terms of content, the discretion is large; in terms of procedure, only intolerable deviation from information properly retrieved constitutes a ground for liability. Hence, managers may in most cases override the machine recommendation. Of course, managers will take into account also factors that are not legal in the narrow sense (legal limits related), namely the preferences of the body of shareholders and other constituencies.

The overall conclusion would seem to be the same as in malpractice cases. Particularly in ML contexts, human oversight, and the possibility to disagree with the model based on entrepreneurial reasons, seems of high importance: ML makes mistakes humans would not make (and vice versa, potentially even more often one day). Hence, in corporate contexts, too, a machine-human-team, which likely works better than either machine or human alone. Importantly, the possibility to disagree often presupposes explainability. Furthermore, withholding individual judgment is not a liability-minimizing strategy for managers, either. Therefore, there are no incentives for refraining from a reasoned departure from the model.

3.2.3 Short summary of case study 2

Summing up, ML models are approaching, and in the future potentially even surpassing, human decision-making capacity also in the corporate mergers realm. However, the legal standard of care should be adjusted only upward by the introduction of such models if (1) they are proven to be consistently outperforming professional actors and other models, and, due to disclosure requirements and multiple

party decision making, when (2) they are used as interactive techniques that can be explained to and critically assessed by humans. Given the size and the timeframe of the transactions, integration into the process is basically not a concern (different from the malpractice cases). As in that case, managers must be obliged to exercise independent judgment and may disagree with the model, based on professional reasons, without risking legal liability.

3.2.4 Key results from case studies 1 and 2: contractual explainability

Both the case study on medical malpractice and on corporate mergers show that, from a legal point of view, explainability is often at least as important as predictive performance to determine whether the law allows or even requires the use of ML tools under professional standards of care. Most notably, this conclusion is independent of the debate on the extent to which the European GDPR does or does not require the explainability of certain algorithmic models (see above, under Sect. 2). The criteria developed for assessing the necessity of explainable AI stem, legally speaking, not from data protection law but from specific domains of contract and tort law. As such, they are not confined to the EU, but apply equally in all countries with negligence regimes, in contract and tort law, similar to our reference jurisdictions: the US and Germany. Since there has been a considerable degree of convergence in EU contract law in the past decades (Twigg-Flessner 2013), the guidelines we develop apply, at least broadly, both in US and in EU law. In the third case study, we therefore examine more in detail the interaction, and trade-off, between performance and explainability from a technical perspective.

4 Explanations and accuracy

As seen, data protection law, but even more importantly contract and tort law may require models to be explainable in the sense of offering an explanation for an automatically made decision. These explanations ensure that algorithms can be introspected to check whether the decisions they made are fair and unbiased. This legal requirement also helps engineers to better understand the models they build. By being able to explain the model's decision, flaws in the training data or the model architecture can be detected, e.g., when the image classifier explains a classification decision by focusing on the background of the image instead of the object to be classified. But what is an explanation that meets the legal requirements and is technically sound?

Miller (2019) investigates explanations from a philosophical, psychological, and cognitive science view and discusses how explanations work for humans. Since 1748, when David Hume presented his ideas on the importance of causality and counterfactual reasoning for explanations (Hume 2016), many new theories have been proposed, but they are all more or less extensions of the idea of counterfactuals. While a causal chain explains a certain state or decision technically, this is typically not accepted as an explanation by the non-expert users of a system. Instead—and these are two major findings of Miller (2019), (1) people

select one or two causes out of possibly infinite causes to be the explanation and (2) explanations are often expected to be contrastive.

Regarding (2), this means that explainable AI can be reduced to finding the difference between two alternatives. Even if the alternative is not explicitly stated, e.g., in image classification, the question of a user could be "why was this image classified as depicting a dog?", seemingly requiring an explanation about the working of the (potentially very complex) classification algorithm, while in reality, the explanation should target a more specific question, e.g.: "Why was this image classified as depicting a dog and not a wolf?". In this case, the explanation could be easier generated by summarizing the differences between dogs and wolves.

Regarding (1), choosing from a multitude of causes to generate an explanation is very context-dependent and potentially influenced by cognitive bias. This selection of causes makes it on the one hand easier to explain decisions of AI algorithms, since not the complete causal chain needs to be presented. On the other hand, selecting the "right" cause accepted by a user with given bias and context is far from trivial. In the past, this was an easier problem: Traditional AI, or more precisely, machine learning (ML) algorithms, were primarily linear models that could be inspected by human experts and explained even to non-professionals. A rule learned by a decision tree model, for example, could state if $x > 3.4$ then $y = 1$. The model explained itself by giving a cause for its decision; no additional explanation was needed.

With more recent and more complex models, such self-explanations became unintuitive or even impossible. With the growing size of complex, non-linear models, not even experts can easily explain the outcome due to the huge number of parameters involved. With the missing inherent explanations, the question arose how to explain the models' output if not even experts can understand the models as a whole. This need for explanation—a basic human desire to learn and trust—shifts the focus away from the model and towards the explanation (Monroe 2018). Just as humans have different explanations for their decisions and behavior, there are also different kinds of explanations for AI models.

When we look at different models, there seems to be a trade-off between model accuracy and explainability of the model—just like the legal discussion presupposed. In Fig. 1 we extended the graphic from DARPA's explainable AI program[3] to visualize the trade-off between explainability and the accuracy of various model types. This general notion of explainability is broken down into more fine-grained definitions of explainability types in the following section.

4.1 Explanation type

Understanding models in detail with millions of parameters is not feasible. What can be hoped for, though, is to find some explanation that satisfies the need for

[3] https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

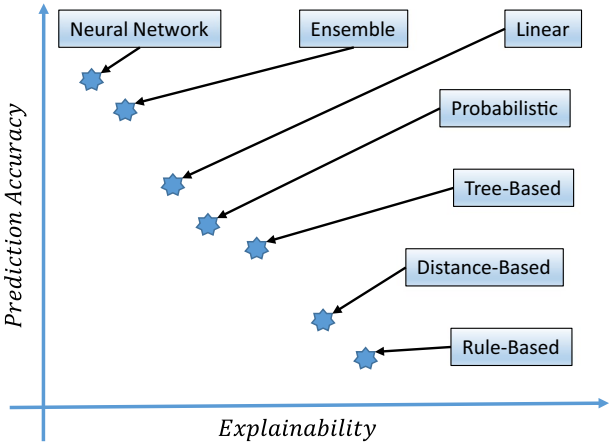

Fig. 1 Accuracy and explainability trade-off

interpretability of models at a specific level of abstraction. Lipton identified different kinds of explanations relevant to machine learning algorithms depending on specific model characteristics (Lipton 2018). Explanations can be broadly associated with *transparency* and post-hoc *interpretability*. Models that are highly transparent are self-explanatory whereas post-hoc interpretability is a characteristic that allows explaining decisions after the model's inference step. The following list is a summary of Lipton (2018).

1. Transparency describes how easily a model can be understood.

   (a) At the level of the entire model: *simulatability*
   (b) At the level of individual components: *decomposability*
   (c) At the level of the learning algorithm: *algorithmic transparency*

2. Post-hoc interpretability describes how easily a decision by a learned model can be explained.

   (a) By visualizing what a model has learned: *visualization*
   (b) By analyzing the parameters for a single decision: *local explanations*
   (c) By finding and presenting the most similar examples: *explanation by example*

Lipton associates explainability with interpretability whereas understanding is connected with transparency. Biran and Cotton (2017) define an explainable model based on how easy it is for a human to understand a model's decision either by introspection or by a generated explanation. Both emphasize the dual character of transparency and interpretability for explanations.

Table 1 Different models and their accessibility to explanations

Models	Transparency		
	Simulatability	Decomposability	Algorithmic
Rule-based	High	High	High
Tree-based	High	High	Medium
Linear	Medium	Medium	High
Distance-based	Medium	Medium	High
Probabilistic	Medium	Medium	High
Ensemble	Low	Low	Medium
Neural network	Low	Low	Low

These different notions of explanations lend themselves to different kinds of machine learning models. In general, less complex models are more transparent and necessitate less often post-hoc interpretations to explain their decisions sufficiently.

4.2 ML model type

Flach (2012) categorizes machine learning models into seven broad classes:

1. *Rule models* learn easy to interpret rules. A prototypical model is the naive Bayes model (NB).
2. *Tree models* learn a hierarchy of rules. A prototypical model is the decision tree model (DT).
3. *Linear models* learn to combine input variables linearly to produce output. A prototypical model is the support vector machine model (SVM).
4. *Distance-based models* learn similarity between instances. A prototypical model is the k-nearest neighbour model (kNN).
5. *Probabilistic models* learn (conditional) probabilities for possible outcomes. A prototypical model is logistic regression (LR).
6. *Ensembles* learn to combine different algorithms. A prototypical model is the random forest model (RF).
7. *Deep neural networks* learn weight tensors to transform an input to a target output. Prototypical models are convolutional neural networks (CNN).

There are also other categorizations to divide the model space and the proposed classification of these models is not the only valid one.[4] Naive Bayes is rather a probabilistic model but it is very simple to deduce rules from it therefor we use naive Bayes instead of, e.g., association rules as the prototypical rule-based model. This categorization shall serve our purpose to investigate explainability in more detail based on individual model types. To this end, we picked a prototypical model for each model class and looked at how accessible it is to each explanation type.

[4] See, e.g., https://www.darpa.mil/attachments/XAIProgramUpdate.pdf for a slightly different grouping.

Table 2 Different models and their accessibility to interpretation and performance

Models	Post-hoc interpretability			Performance
	Visualization	Local	By example	
Rule-based	High	High	Medium	Low
Tree-based	High	High	Medium	Medium
Linear	High	High	High	Medium
Distance-based	Low	Medium	High	Low
Probabilistic	High	High	High	Medium
Ensemble	Low	Medium	Medium	High
Deep neural network	High	Medium	High	High

4.3 Model/explanation matrix

We investigate how well individual models are suited for the different types of explanations. To this end, we group the models for each explanation type into *high*, *medium*, and *low*. "High" means that the model exhibits a specific property to a large degree, and "low" means that the model possesses this property only to a small extent. Tables 1 and 2 shows our classification of the suitability of the various model types.

Regarding the results, there is an apparent trade-off between transparency and performance. Especially for complex problems, such as object recognition in computer vision or machine translation in natural language processing, the best performing models are also the ones inherently not transparent. Approaches for explaining these models are subject to current research (Samek et al. 2017; Montavon et al. 2017; Bach et al. 2015; Lapuschkin et al. 2019; Rudin 2019). For highly complex models visualization seems to be the most promising type of explanation. The main idea researchers are pursuing to this end is to visualize salient parts of the input, e.g., of an image (Simonyan et al. 2013), or of text (Arras et al. 2017). Gilpin et al. (2018) argue that these current methods produce explanations that are not sufficient, and that combining these methods and measures is needed to obtain explanations from deep learning models that are acceptable as such for humans.

4.4 Example: automatic spam detection

To substantiate our discussion above, we present our evaluation of a concrete example use-case. A standard task for machine learning is document classification. To make the problem as clear and simple as possible, we chose spam classification, i.e., a binary classification of a given SMS into "spam" or "no-spam". We use a dataset of 5574 SMS messages (Almeida et al. 2013). For simplicity, we ignore additional information, such as the sender of an SMS, and use only the actual text in the message. To compare the models, we split the data into training and test set. In the training set, there are 3233 ham SMS and 500 spam SMS. In the test set there are 1592 ham SMS and 247 spam SMS.

Table 3 Different models and their classification accuracy for SMS-spam prediction

Models	Classification accuracy
Deep neural network: CNN	0.985
Linear: SVM	0.983
Probabilistic: LR	0.978
Rule-based: NB	0.977
Ensemble: RF	0.976
Tree-based: DT	0.965
Distance-based: kNN	0.948

Spam detection does not require overly complicated models. The existence of certain keywords (e.g., "promotion", "offer", "Viagra", etc.) is often enough to identify unsolicited messages. Given this simple task where a deeper understanding of sentences or context is not necessary, simpler models not only explain the predictions better but even outperform more complex models. Our results for SMS-spam classification, which are similar to the ones reported by Almeida et al. (2013), are shown in Table 3. We compare the prototypical models as described in Sect. 4.2. For this task, very simple models, such as naive Bayes can outperform more complex models, e.g. ensemble models, such as random forests, under certain conditions, e.g., if only a small amount of annotated training data is available. These simple models can easily explain their predictions. Since they usually assume that input features are independent (naive Bayes assumption), the mere occurrence of a word in the message explains a certain prediction. E.g., the word "free" (lowercase) occurs 52 times in spam SMS in comparison to 43 times in ham SMS, even though there are much more ham SMS than spam ones. If one considers the word "FREE" (uppercase), the ratio is overwhelmingly in favor of spam: it occurs only once in a ham SMS, but 134 times in spam SMS. A simple rule, such as *if an SMS contains the terms "free" and "camcorder" then classify it as spam*, results in 100% accuracy without any wrong predictions in the complete data set and directly explains the decision (occurrence of these two words).

Therefore, it is not enough from a computer science research perspective to focus on making complex models explainable, but also to choose the right model for the task at hand. The trade-off between performance and accuracy might only hold for a subset of particularly complex problems (Rudin 2019; Chen et al. 2018), such as image classification and machine translation. Unfortunately, the human need for explanations is higher for complex problems, where models can achieve supra-human accuracy. "Obvious" predictions do not require explanation.

4.5 Outlook: interactions between technology and the law

Different models differ widely in the degree to which the decisive features and weights of the model can be extracted (Burrell 2016). Under decision trees, at least in theory, the algorithmic output can be determined by walking from the trees' root

along branches to the terminal leaves. At the very opposite end of the spectrum, deep neural networks do not lend themselves to an easy explanation, both because it is difficult to extract decision weights from the highly non-linear mathematical model (Li et al. 2018), and because hidden layers may operate with features that bear no resemblance to any concepts easily understandable by humans (Burrell 2016). However, recent advances in interpretable machine learning show that it may be possible to train a second model to approximate, if not globally, at least locally the behavior of the model that needs to be explained (Ribeiro et al. 2016). This means that, at least for specific instances of decisions, positive and negative decision factors can potentially be identified. What is more, other approaches even allow to test to what extent specific, human-understandable concepts were used by the deep neural network in its classification process (Kim et al. 2018). Importantly, this may allow a doctor, for example, to determine whether the algorithmic output is likely based on an "accurate understanding" of the training data; this helps to evaluate whether the performance of the model translates to real-world examples, or not (Ribeiro et al. 2016; Kim et al. 2018; Lapuschkin et al. 2019). Similarly, risk assessment experts in companies could seek to verify, via interpretable machine learning, to what extent the model fits the specific context they are deploying it in. As has been noted, more accurate models [refined via LASSO, for example Tibshirani (1996)] may even sometimes be more explainable as they pick up less spurious correlations from the training data and focus on the "relevant" correlations (Selbst and Barocas 2018); these can then be the basis for an explanation. All in all, while interpretable machine learning certainly still has a long way to go, the advent of explainability requirements will likely inject even more force into these models. Their development and deployment will likely benefit both decision makers and data subjects in the long run. Again, the law impels the adoption of socially desirable machine learning technology.

Conversely, technological progress has important implications for the law. In order to make the cost-benefit analysis trading off the accuracy against the explainability of different models, non-technical professionals like medical doctors and corporate managers will likely either need to acquire technical expertise on their own or work with professional developers of machine learning applications. This is particularly important since, as our experiment with the spam prediction model showed, increased explainability need not necessarily imply reduced accuracy. To be sure, we selected a simple problem (spam classification) to show the trade-off between accuracy and simplicity/transparency of models. For more complicated problem settings, general explanation generation algorithms and trade-offs between accuracy and explainability are hard to come by. Hence, often, there are only very specialized solutions for experts to inspect the models, making the transdisciplinary teaming even more important.

Nevertheless, data-driven validation of the trade-off in the respective areas of application seems necessary. In fact, the increased adoption of ML tools will lead, we posit, not only to an increased importance of human-machine-teaming, but also to the emergence of more interdisciplinary work units within business organizations. In order to meet the requirements imposed by contract and tort law, sector specific experts, such as medical doctors or corporate managers, will often have to team up

with professional developers of machine learning applications in order to correctly evaluate the different models at their disposal. This is certainly not a bad thing: potentially, the law in this sense sets incentives for fruitful interdisciplinary collaborations that lead to progress both on the professional and on the technical side of the respective fields.

5 Conclusion

In conclusion, we have shown that the relationship of AI and the law is more complex than policy analysis often suggests. The law not only limits AI technology; it often sets incentives for, or even mandates the application of, the use of models when their very use minimizes the risk of liability. However, the devil is in the details. Machine learning research offers an ever wider array of input patterns, model types, and different performance metrics. Some of these will be particularly suited to certain legal tasks, such as risk analysis or contract compliance; others will indeed face limitations from anti-discrimination or data protection law. Even these constraints, however, may eventually be helpful for guiding machine learning research, by setting incentives for developing socially optimal, accurate and interpretable models—or at least for raising awareness among professionals of the extent to which these parameters have to be traded off to make an informed decision about whether or not to adopt a specific ML model in the fist place.

Acknowledgements : Open Access funding provided by Projekt DEAL.

Funding The research for this article was in part (Philipp Hacker) supported by an AXA Postdoctoral Scholarship awarded by the AXA Research Fund.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Almeida T, Hidalgo JMG, Silva TP (2013) Towards sms spam filtering: results under a new dataset. Int J Inf Secur Sci 2(1):1–18

Arras L, Horn F, Montavon G, Müller KR, Samek W (2017) What is relevant in a text document?: An interpretable machine learning approach. PloS one 12(8):e0181142

Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one 10(7):e0130140
Biran O, Cotton C (2017) Explanation and justification in machine learning: a survey. In: IJCAI-17 workshop on explainable AI (XAI), pp 1–6
Burrell J (2016) How the machine thinks: understanding opacity in machine learning algorithms. Big Data Soc 3(1):2053951715622512
Calo R (2016) Robots in American law. University of Washington School of Law Research Paper (2016-04)
Caruana R (2015) Intelligible models for healthcare. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, vol 21, pp 1721–1730
Chen C, Lin K, Rudin C, Shaposhnik Y, Wang S, Wang T (2018) An interpretable model with globally consistent explanations for credit risk. arXiv preprint arXiv:181112615
Cowgill B (2017) Automating judgement and decision-making: theory and evidence from résumé screening. In: Columbia University, 2015 Empirical Management Conference
De Fauw J (2018) Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 24(9):1342
Ding Y (2018) A deep learning model to predict a diagnosis of alzheimer disease by using 18F-FDG PET of the brain. Radiology 290:456–464
Doshi-Velez F (2017) Accountability of AI under the law: the role of explanation. arXiv preprint arXiv:171101134
Droste W (2018) Intelligente Medizinprodukte. Zeitschrift fr das gesamte Medizinprodukterecht 18(4):109–114
Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115
Flach P (2012) Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press, Cambridge
Froomkin M (2018) When AIs outperform doctors: confronting the challenges of a tort-induced overreliance on machine learning. Arizona Law Review, Forthcoming; University of Miami Legal Studies 18(3)
Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an overview of interpretability of machine learning. In: 5th International conference on data science and advanced analytics (DSAA), pp 80–89
Goodman B, Flaxman S (2016) Eu regulations on algorithmic decision-making and a right to explanation. In: ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. arxiv:1606.08813 v1
Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a right to explanation. AI Mag 38(3):50–57
Greenberg MD (2009) Medical malpractice and new devices: defining an elusive standard of care. Health Matrix 19:423
Grundmann S (2012) European company law—organization finance and capital markets. Intersentia, Cambridge
Hart D (2000) Evidenz-basierte Medizin und Gesundheitsrecht. MedR-Medizinrecht 18(1):1–5
Hopt KJ, Roth WH (2015) Sorgfaltspflicht und verantwortlichkeit der vorstandsmitglieder. In: Grokommentar Aktiengesetz, vol 4(2). de Gruyter, p §93
Hume D (2016) An enquiry concerning human understanding. In: Seven masterpieces of philosophy, Routledge, pp 191–284
Jaklevic C (2017) MD Anderson cancer centers IBM Watson project fails, and so did the journalism related to it. Health News Review
Jiang T (2018) Using machine learning to analyze merger activity. Working paper UC Davis, 16 March 2018. http://giovanniperi.ucdavis.edu/uploads/5/6/8/2/56826033/tiffany_jiang.pdf. Accessed 10 Nov 2019
Katzenmeier C (2006) Aufklärung über neue medizinische Behandlungsmethoden—Robodoc
Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F (2018) Interpretability beyond feature attribution: quantitative testing with concept activation vectors (tcav). In: International conference on machine learning, pp 2673–2682
Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA (2018) The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 24(11):1716

Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI (2015) Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13:8–17
Kraakman R (2017) The anatomy of corporate law—a comparative and functional approach. Oxford University Press, Oxford
Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR (2019) Unmasking clever hans predictors and assessing what machines really learn. Nat Commun 10(1):1096
Laufs A (1990) Die Entwicklung des Arztrechts. NJur Wochenschr 24:1507–1513
Li K (2018) Corporate culture and mergers and acquisitions: evidence from machine learning. https://haslam.utk.edu/sites/default/files/files/LiMaiShenYan_Corporate%20Culture%20and%20Mergers%20and%20Acquisitions_20180429.pdf. Accessed 10 Nov 2019
Li O, Liu H, Chen C, Rudin C (2018) Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In: Thirty-second AAAI conference on artificial intelligence
Lipton ZC (2018) The mythos of model interpretability. Queue 16(3):30:31–30:57 10.1145/3236386.3241340
Malgieri G, Comandé G (2017) Why a right to legibility of automated decision-making exists in the general data protection regulation. Int Data Priv Law 7(6):243–265
Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38
Monroe D (2018) AI, explain yourself. Commun ACM 61(11):11–13
Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR (2017) Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognit 65:211–222
Mori T, Uchihira N (2018) Balancing the trade-off between accuracy and interpretability in software defect prediction. Empirical Softw Eng 24:1–47
Party ADPW (2017) Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679
Press X (2018) China focus: AI beats human doctors in neuroimaging recognition contest. http://www.xinhuanet.com/english/2018-06/30/c_137292451.htm. Accessed 10 Nov 2019
Reed C, Kennedy E, Silva S (2016) Responsibility, autonomy and accountability: legal liability for machine learning. Queen Mary School of Law Legal Studies Research Paper (243)
Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you?: explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1135–1144
Rosoff AJ (2001) Evidence-based medicine and the law: the courts confront clinical practice guidelines. J Health Politics Policy Law 26(2):327–368
Ross C, Swetlitz I (2018) IBMs watson supercomputer recommended unsafe and incorrect cancer treatments, internal documents show. Stat News. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafeincorrect-treatments. Accessed 10 Nov 2019
Royal Society (2017) Machine learning: the power and promise of computers that learn by example
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206
Rudin C, Ustun B (2018) Optimized scoring systems: toward trust in machine learning for healthcare and criminal justice. Interfaces 48(5):449–466
Samek W, Wiegand T, Müller KR (2017) Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:170808296
Selbst AD (2019) Negligence and AI's human users. Boston University Law Review, Forthcoming
Selbst AD, Barocas S (2018) The intuitive appeal of explainable machines. Fordham L Rev 87:1085
Selbst AD, Powles J (2017) Meaningful information and the right to explanation. Int Data Priv Law 7(4):233–242
Shao B, Asatani K, Sataka I (2018) Analyzing mergers and acquisitions (M&A) in Japan using AI methods. In: Proceedings of the annual conference of JSAI. The Japanese Society for Artificial Intelligence, pp 2J1–214
Shibayama S (2008) Effect of mergers and acquisitions on drug discovery: perspective from a case study of a Japanese pharmaceutical company. Drug Discov Today 13:86–93
Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps

Thomas S (2017) Artificial intelligence, medical malpractice, and the end of defensive medicine. https://blog.petrieflom.law.harvard.edu/2017/01/26/artificial-intelligence-medical-malpractice-and-the-end-of-defensive-medicine/. Accessed 10 Nov 2019

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol) 58:267–288

Timmermans S, Mauck A (2005) The promises and pitfalls of evidence-based medicine. Health Aff 24(1):18–28

Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25(1):44

Twigg-Flessner C (2013) The Europeanisation of European contract law. Routledge, Abingdon

US Food and Drug Administration (2018) FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. News Release, April

Varzaly J (2012) Protecting the authority of directors: an empirical analysis of the statutory business judgment rule. J Corp Law Stud 12:429–463

Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Priv Law 7(2):76–99

Wagner G (2018) Comment on §630a BGB. In: Münchener Kommentar zum BGB, Verlag CH Beck, pp 405–418

Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data? PloS one 12(4):e0174944

Wischmeyer T (2018) Regulierung intelligenter systeme. Archiv des öffentlichen Rechts 143(1):1–66

Witten IH, Frank E, Hall MA, Pal CJ (2016) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington

Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 15(11):e1002683

Zuo Z, Loster M, Krestel R, Naumann F (2017) Uncovering business relationships: context-sensitive relationship extraction for difficult relationship types. In: Proceedings of the conference Lernen, Wissen, Daten, Analysen (LWDA), vol 1917, pp 271–283