Explainable AI and Law

An Evidential Survey

McGregor Richmond, Karen; Muddamsetty, Satya M.; Gammeltoft-Hansen, Thomas; Palmer Olsen, Henrik; Moeslund, Thomas B.

Published in:
Digital Society: Ethics, Socio-Legal and Governance of Digital Technology

DOI (link to publication from Publisher):
10.1007/s44206-023-00081-z

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Digital Society (2024) 3:1
https://doi.org/10.1007/s44206-023-00081-z

ORIGINAL PAPER

Explainable AI and Law: An Evidential Survey

Karen McGregor Richmond[1] · **Satya M. Muddamsetty**[2] · **Thomas Gammeltoft-Hansen**[1] · **Henrik Palmer Olsen**[1] · **Thomas B. Moeslund**[2]

Received: 11 June 2023 / Accepted: 31 October 2023 / Published online: 19 December 2023

Abstract

Decisions made by legal adjudicators and administrative decision-makers often found upon a reservoir of stored experiences, from which is drawn a tacit body of expert knowledge. Such expertise may be implicit and opaque, even to the decision-makers themselves, and generates obstacles when implementing AI for automated decision-making tasks within the legal field, since, to the extent that AI-powered decision-making tools must found upon a stock of domain expertise, opacities may proliferate. This raises particular issues within the legal domain, which requires a high level of accountability, thus transparency. This requires enhanced explainability, which entails that a heterogeneous body of stakeholders understand the mechanism underlying the algorithm to the extent that an explanation can be furnished. However, the "black-box" nature of some AI variants, such as deep learning, remains unresolved, and many machine decisions therefore remain poorly understood. This survey paper, based upon a unique interdisciplinary collaboration between legal and AI experts, provides a review of the explainability spectrum, as informed by a systematic survey of relevant research papers, and categorises the results. The article establishes a novel taxonomy, linking the differing forms of legal inference at play within particular legal sub-domains to specific forms of algorithmic decision-making. The diverse categories demonstrate different dimensions in explainable AI (XAI) research. Thus, the survey departs from the preceding monolithic approach to legal reasoning and decision-making by incorporating heterogeneity in legal logics: a feature which requires elaboration, and should be accounted for when designing AI-driven decision-making systems for the legal field. It is thereby hoped that administrative decision-makers, court adjudicators, researchers, and practitioners can gain unique insights into explainability, and utilise the survey as the basis for further research within the field.

This research is funded by the Danish National Research Foundation Grant no. DNRF169 and conducted under the auspices of the Danish National Research Foundation's Centre of Excellence for Global Mobility Law.

Extended author information available on the last page of the article

Keywords XAI · Legal reasoning · Legal logics · Explainability · Artificial intelligence · Evidence

1 Introduction

In conceptual terms, the legal domain can be thought of as a database of statutory rules, the administrative and judicial decisions shaped by these rules, in addition to a body of meta-rules encompassing theories, procedures, and hierarchies of authority, all of which are continuously updated and made enforceable through institutions (government, courts, administrative tribunals, etc.), collectively designated with the task of regulating behaviour through transparent and rational legal decision-making. Whilst the legal domain is governed by human agents, recent developments in artificial intelligence (AI) algorithms suggest that the technology is — or may soon be — able to augment, or even replace, administrative decision-making and legal adjudication across a number of diverse legal fields. Attempts to implement AI within the legal field have a long pedigree, and legal implementation of AI technology remains a highly active field of research. Much of this research aims to render the law more comprehensible, manageable, useful, accessible, or even predictable, in terms of its outputs (Surden, 2019). Indeed, the legal domain possesses a number of characteristics which render it amenable to AI-driven developments, including, but not limited to: engagement with diverse categories of knowledge; large numbers of similar decisions (in certain fields of law); explicit styles, and standards of justification; different modalities of reasoning; specialised repositories of knowledge; and a variety of task orientations that together make legal AI an especially interesting and challenging target for developers (Rissland et al., 2003).

The attempt to capture, and replicate, the modalities of legal reasoning has therefore proven itself to be an enduring objective of AI application (Rissland et al., 2003). Research in the AI and Law domain has focused on modelling both the factual basis underpinning legal decision-making, in addition to the nexus between legal facts and the rules of law, the latter understood as the aggregate of "legal sources", including both legislation and case law (i.e. previous decisions made in similar cases). Within the legal domain, however, there is significant divergence in terms of how legal reasoning is implemented, a phenomenon not hitherto addressed in the literature. The specific affordances of legal reasoning are dependent upon the overarching legal system, in tandem with the area of specialisation in question. Thus, each category of legal specialisation, e.g. criminal, administrative, or private law, is characterised by different legal institutions, a different evidentiary threshold, and different procedural rules. Such functional divergence generates divergence in modes of legal reasoning, from rule-based inductive argumentation, to narrative-based holistic approaches. Therefore, in their reasoning processes, criminal courts differ markedly from administrative tribunals [49]. Secondly, these elements can vary significantly across countries, most notably between jurisdictions adhering to the common law (most notably the USA, England, Canada, and Australia), the civil law (e.g. France, Germany,

Italy), or even a mixed common/civil law system (e.g. Scotland, South Africa, Israel), whose traditions in turn ascribe differing values to, for example, judicial inquiry or rules derived from prior cases. Each of these elements — the field of law, the legal system, the procedural setup of the decision-making process, the legal institutions, and secondary rules pertaining to decision-making procedure — collectively impact on the utilisation of AI to better support, augment, or automate legal decision-making. Consequently, as practical AI applications in law expand, and are exported across different fields of law and jurisdictions, there is an inherent risk that systems trained or developed in one area of law, or one jurisdiction, characterised by a particular mode of legal reasoning, may be blind to the differing structures, or contexts, of putatively similar domains in which they are later applied.

This risk is compounded by the direction of current technological advances in AI modelling, and the difficulties of explaining individual model outcomes to legal professionals. Traditionally, symbolic AI approaches have been used for modelling legal reasoning. Symbolic AI is based upon rules created by human agents. Thus, symbolic AI differs from machine learning, whereby an algorithm learns rules as it establishes correlations between inputs and outputs. TAXMAN (McCarty, 1976) was the first system to implement this form of legal reasoning model, utilising a theorem-proving approach to corporate tax law issues. It was succeeded by HYPO (Ashley, 1989), CATO (Aleven, 1997), and IBP, or Issue-Based Prediction (Bourcier, 2003). All of these systems operated within highly formalised and tightly delineated areas of the law. However, attempts to deal with more complex legal domains, and areas of law where rules are more likely to conflict, or be subject to regular change, require comparatively advanced methodologies; hence, more complex systems are required to analyse, and represent, legal reasoning and legal knowledge in such domains. In recent years, a stream of research has thus been directed towards the prediction of case outcomes using algorithms applied to large legal datasets (Aletras et al., 2016; Chen & Eagel, 2017). These efforts are founded upon machine learning, as opposed to symbolic AI, and therefore use an algorithm to establish rules from correlations that emerge within large datasets. Concerns have been raised in relation to the development, deployment, updating, and interpretability of such models (Górski et al., 2021). Nonetheless, concerns over explainability have not hampered the rise of highly opaque decision-making systems, such as Deep Neural Networks (DNNs) which have experienced remarkable growth, including in the legal domain, where DNNs have been used to extract legal information by learning from example texts (Chalkidis & Kampas, 2019; Lippi et al., 2019). Furthermore, it follows that explainability, legally situated in the duty to ensure that decisions are capable of explanation, becomes more difficult to compass when models are applied across different legal domains and jurisdictions.

Similarly, the newer generation of legal AI models often demonstrates robust performance, yet their outputs remain comparatively unclear and "black-boxed" due to the enormous "parameter space" at play within the systems. This leads to a significant problem when applied in the legal field, where applications of AI require that outcomes are justifiable (Brożek et al., 2023) and generated in conformity with pre-established and identifiable norms and legal reasoning (Antoniou et al., 2022).

Indeed, lack of suitable explanation has been a significant barrier to the adoption of such technologies by both public bodies and private law firms. Despite attempts to develop practical solutions (see, for example, Branting, 2003), significant problems remain, which retard the development of fully explainable AI to the legal arena (Bibal et al., 2021). Given the rapid developments in AI and Law, and the growing gap between performance and explainability, the struggle to ensure that AI systems remain trustworthy, rationally explicable, and ethically defensible has thus emerged as a key issue.

One response to this problem is the turn to Explainable AI (XAI), an emerging body of research specifically focused on understanding how AI systems make decisions and how to make these decisions more interpretable, and explainable, to humans (Brożek et al., 2023). In the legal domain, applications of XAI still remain limited, but proponents have argued that XAI can aid judges who want to rely on algorithms for decision support, litigants who want to persuade judges that their use of algorithms is lawful, and defendants who want to question AI-based administrative decisions (Palmirani et al., 2012). The turn to XAI in itself, however, does not necessarily eclipse the above-mentioned critiques of AI applications in law since, as has been postulated, XAI models may run the risk of propagating "translation errors" if applied indiscriminately across substantively variegated legal systems and sub-fields. Further, it has been postulated that the development of methods for explaining black box AI should be eschewed in favour of creating models that are interpretable ab initio through their transparent design. Further, that the focus on post facto explanation is likely to perpetuate bad practice and may potentially propagate harm (Rudin, 2019).

Such caveats should be borne in mind when compassing the present article, which offers a novel approach to surveying explainability — understood as the ability to provide a legally cogent explanation — in AI and Law research, through paying particular attention to the underlying legal reasoning processes and procedures upon which the research articles are founded. By unpacking the current state of the art, in light of the different legal sub-domains and procedures that the research relates to, this survey seeks to illuminate the potential problems generated by indiscriminately applying algorithmic models to legal administrative decision-making (such as that encountered in the asylum domain), and courtroom adjudication. Further, it aims to highlight the benefits of applying appropriate AI models to particular legal sub-fields and modes of legal reasoning, whilst correctly situating these algorithmic models as a product of interdisciplinary research.

The present article thus seeks to address limitations in previously published surveys (Alikhademi et al., 2022; Atkinson & Bench-Capon, 2019), centrally the failure to account for the heterogeneous nature of reasoning, adjudication, and administrative decision-making across legal sub-domains. To reiterate, in recent years a significant number of papers have concerned themselves with explainability and legal reasoning processes as part of the general turn to AI. This remains the focus of the instant survey. However, whilst previous overviews have been conducted both at the general level (Evans et al., 2022; Guidotti et al., 2018; Islam et al., 2022; Schwalbe & Finzel, 2023; Tjoa & Guan, 2020; Zhang et al., 2022) and, in connection with specific sub-domains (Alikhademi et al., 2022; Atkinson & Bench-Capon, 2019;

Matulionyte & Hanif, 2021), up-to-date systematic reviews on AI and Law — particularly those which are alive to the heterogenous nature of legal reasoning — remain scarce. Thus, in an attempt to address the lacuna, the present article provides a systematic survey reaching across a range of legal sub-domains and, on this basis, sets out a legally coherent systematisation of the field. In so doing, the authors take inspiration from the taxonomy of explanation strategies in AI and Law developed by Atkinson et al. (2020), but further develop their work to bring attention to the link between different forms of legal procedure and legal reasoning, on the one hand, and the variegated models of AI, on the other. The main aims and contributions of the article are therefore summarised as follows:

1. Based on a systematic survey, the article establishes a novel taxonomy, linking the differing forms of legal inference, adjudication, and administrative decision-making to specific forms of algorithmic decision-making.
2. The paper provides an interdisciplinary discussion — encompassing both legal and technical perspectives — to develop understandings of the intrinsic relationship between differing modes of legal reasoning, in light of algorithmic developments, evaluation measures, and challenges.
3. Based upon the survey, the article outlines a number of key XAI challenges within the legal domain, which have yet to be adequately addressed. Specifically, the authors identify ongoing ambiguity around the concepts and metrics used to evaluate the explainability of AI models.
4. The paper outlines future research directions, in particular the necessity for further research addressing the need to make DNN models more transparent and understandable by actors in the legal domain.

The remainder of the survey paper is organised as follows. The conceptual background is explained in Section 2. Section 3 presents the survey methodology. In Section 4, a taxonomy of commonly encountered XAI categories is presented. Lastly, the future research directions of the findings of this study, and its conclusions, are presented in Section 5.

2 Conceptual Background: Explainability in AI and Explanations in Law

In order to proceed with the instant literature survey and analysis, it is necessary to first establish a common point of understanding on what the term explainability means, both in the context of AI and within the legal domain. Notably, the instant survey uncovered no commonly accepted definition of this term within either discipline. Rather, explainability is used to signify a variety of different phenomena. This section discusses the differences — and potential trade-offs — between explanatory demands in the context of AI development, and explainability as understood in the legal domain. The sophistication of AI-powered systems has lately increased to such an extent that almost no human intervention

is required for their design and deployment. However, when decisions derived from these ubiquitous AI systems give rise to significant social effects (as when applied to medicine, law, defence, and other sensitive domains) there is a countervailing requirement to comprehend how AI algorithms arrive at their decisions (Burrell, 2016). It is notable that early AI systems were comparatively interpretable, since the majority of them used logic-based symbolic AI. These models were inherently transparent and explainable, having been derived from a set of human-generated rules. However, over the course of the past 20 years, the shift from symbolic AI to machine learning has led to the rise of more opaque decision-making systems. These models are fundamentally "black-boxed", the method by which the algorithm arrives at its decisions having been based on correlations between features which are not logically derived, and frequently not apprehensible to humans. This is especially so with models utilising Deep Neural Networks (DNNs) (Barredo et al., 2020). In law, the inability to apprehend the reasoning behind the algorithm's recommendations can significantly disadvantage those affected by the recommendations. Furthermore, opaque models can undermine citizen's sense of fairness and trust, particularly when used by the government, legal firms, and in the field of criminal justice, where these can undermine a defendant's ability to present a cohesive defence (Deeks, 2019). As a result of these issues, a new research field — explainability in AI, also known as Explainable AI (XAI) — has emerged, which seeks to address the opacity of black-box AI models, and provide human-understandable explanations. Explanation itself is a malleable concept, however, and has been studied extensively by philosophers, particularly by researchers in the philosophy of science and those working in the field of science and technology studies (STS) (Sørmo et al., 2005). In the context of AI, explanations are generally interpreted in two different ways (Aamodt, 1991). One interpretation deals with explanation as part of the reasoning process itself, and the other interpretation deals with usage, and with functional aspects, attempting to make the reasoning process, its output, or the application of the result understandable to the user. Further, providing explanations relative to AI models is guided by four principle functions:

1. An AI system should supply evidence, support, or reasoning for each output.
2. An AI system should provide explanations that its users can understand.
3. Explanation accuracy. An explanation should accurately reflect the process the AI system used to arrive at the output.
4. Knowledge limits. An AI system should operate only under the conditions it was designed for, and refrain from providing an output when it lacks sufficient confidence in the result.

Applied within the legal field, the principles are modified by the requirement to demonstrate that the system operates in conformity with overarching procedural regulations, most notably that pertaining to General Purpose AI Systems (GPAIS) (Council of the European Union, 2021; Gutierrez et al., 2023), in addition to norms of rationality and appropriateness (Rosengrün, 2022). Across the

majority of developed legal jurisdictions, decision-makers are required to provide a reasoned explanation for an administrative or judicial decision (though juries are not required to provide reasons). Such explanations serve as justification for the exercise of authority and discretion that is instantiated by the decision itself. In contrast to such legal explanations, which aim at justification, AI explainability aims at technical comprehension: it refers to the ability to explain the inner workings of a system (Barredo et al., 2020). Hence, a difference in principle has developed between legal explanations (justification of decision-outcome) and AI explanation (causal understanding of how the AI produced its decision). It is important to keep this in mind, and not to confuse the two kinds of explanation. For example, a legal standard for the explanation (justification) of administrative decision-making exists across all main jurisdictions in Europe. But this standard of justification is not one that demands an exhaustive causal explanation of the decision outcome. The authors discovered, when surveying a diverse sample of national jurisdictions (Germany, France, Denmark, and the UK) and regional frameworks (EU law and European Human Rights law), that whilst explanatory requirements differ slightly amongst these jurisdictions, they still conform to a general principle that does not extend into causal explainability (Palmer & Cohen, 2022). Legal explainability is, in other words, not coterminous with technical explainability and vice versa, though it may be of central importance for a legal decision-maker to apprehend the technical features of an automated decision-making model. Further, in the context of research on the use of AI to support, augment, or automate legal decision-making, there is good reason to focus on the interaction of these two forms of explanation. Taking into account the needs of data scientists, explainability is important in order to enhance system robustness (Longo et al., 2020), and to enable more accurate diagnostic analysis, which can in turn help to forestall bias, unfairness, and discrimination (Mehrabi et al., 2021; Wachter et al., 2021). Turning from field-specific differences with regard to explainability, the surveyed literature was notable for methodological divergence. Indeed, XAI methods may be categorised into three major classes: intrinsic explainability (Belle & Papantonis, 2021), post hoc explanations (Barredo et al., 2020), and example-based explanation methods (Adadi & Berrada, 2018). The resulting taxonomy of XAI methods is shown below in Fig. 1. An intrinsically explainable method is typified by the fact that users can understand the decision-making process, or the basis of the technique, without additional information. Typical intrinsic methods include linear regression, logistic regression, *k*-nearest neighbour, rule-based learners, general additive models, Bayesian models, and decision trees (Burrell, 2016). Second, post hoc XAI methods approximate the behaviour of an algorithmic model by extracting relationships between feature values and predictions (Barredo et al., 2020). Thus, they are able to achieve explainability of opaque models without sacrificing system performance. Examples of post hoc explainability methods include attention mechanism, text explanation, visual explanation, local explanation, explanation by simplification, and feature relevance (Belle & Papantonis, 2021). Post hoc methods based upon feature relevance and visual explanations also generate sub-classes, including Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016), Shapley

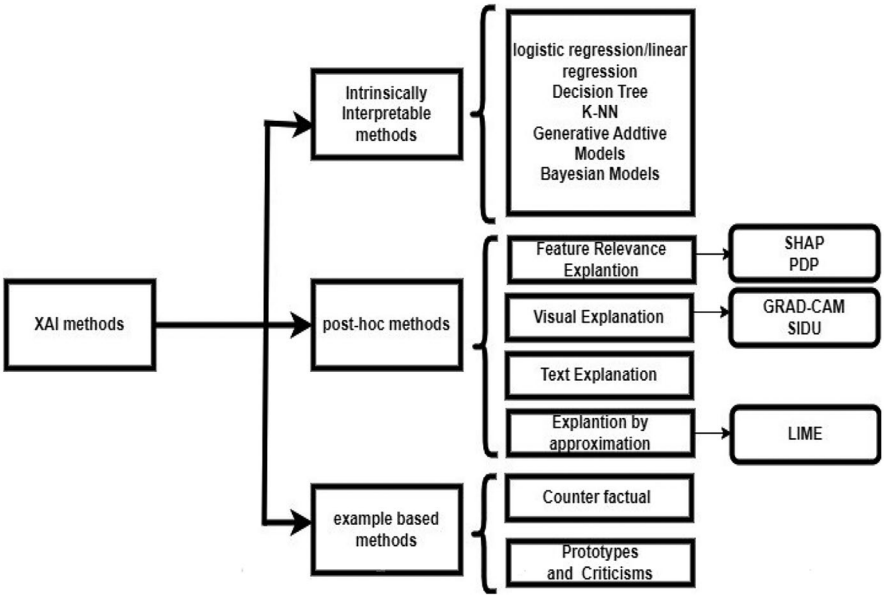

Fig. 1 A categorisation of the principal state-of-the art XAI methods encountered in the literature

Additive explanations (SHAP) (Lundberg & Lee, 2017), class activation mapping (CAM) (Zhou et al., 2016), Gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017), and Similarity Difference and Uniqueness method (SIDU) (Muddamsetty et al., 2022). Third, example-based explanation techniques select particular instances of the data set to explain the behaviour of the machine learning model. The difference between this, and post hoc methods, is that the example-based explanation methods interpret a model by selecting instances of the data set and not by acting on features or transforming the model (Adadi & Berrada, 2018).

So far, the discussion has considered field-specific modes of explainability and divergent categories of explanatory methods. A further theme which emerged from the review was the trade-off between explainability (or interpretability) and performance. Interpretability-versus-performance is a common and long-standing theme, based around the proposition that performance may be measured as a function of accuracy, and that the two are inversely related. High levels of interpretability are achieved at the cost of accuracy (and vice versa). From Fig. 2, it may be further observed that the very first models, such as rule-based models, logistic regressions, and decision trees, were intrinsically interpretable, enjoying a high degree of explainability, but marked by low performance. In contrast, more recent models, based on deep learning, are highly accurate though less interpretable. Future XAI models should ideally aim for high interpretability alongside high performance. However, for the moment, the trade-off between explainability and accuracy remains.

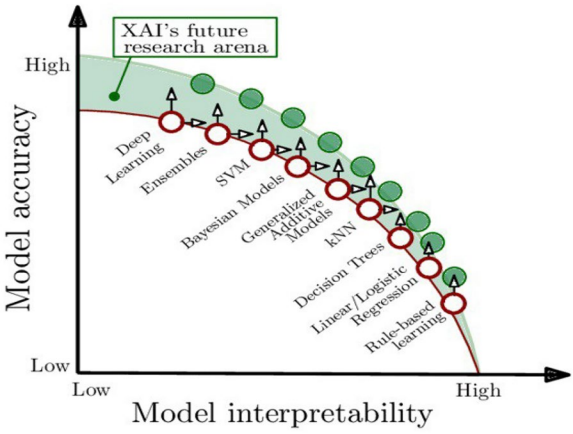

Fig. 2 The trade-off between explainability and performance, adapted from Barredo et al. (2020)

The above offers a common point of departure for the substantive review. In summary, the need for explainability developed alongside a shift from symbolic AI to machine learning. Explainability is moreover subject to differing interpretations across the disciplines, linked to either the reasoning process or usage/functional aspects. More specifically, the four general principles for explainability typically forwarded in AI (justification, understandability, accuracy, and suitability) are modified by the legal domain requirements to demonstrate that the system operates not only in conformity with overarching procedural regulations, but to norms of rationality and appropriateness. From a legal vantage point, explanation thus relates primarily to justification, whereas in AI explainability also aims at technical comprehension [30]. Hence, a difference in principle has developed between legal explanations (justification of decision-outcome) and AI explanation (causal understanding of how the AI produced its decision). The surveyed literature was further notable for methodological divergence. Indeed, XAI methods, it was shown, may be resolved into three major classes: intrinsic explainability (Belle & Papantonis, 2021), post hoc explanations (Barredo et al., 2020), and example-based explanation methods. Lastly, a theme which emerged from the review was the trade-off between explainability (or interpretability) and performance. Having laid out the most salient dimensions of explainability, the subsequent section presents the survey methodology, before discussing how different approaches to explainability contribute to understanding and systematising the surveyed literature.

3 Survey Methodology

In order to operationalise the survey methodology, the authors followed the Preferred Reporting Items for Systematic Literature Reviews and Meta-Analyses guidelines (PRISMA) (Moher et al., 2009). PRISMA is an evidence-based guideline comprising a minimum set of items for reporting systematic reviews and meta-analyses. The guidelines contain clear and robust steps for identifying and

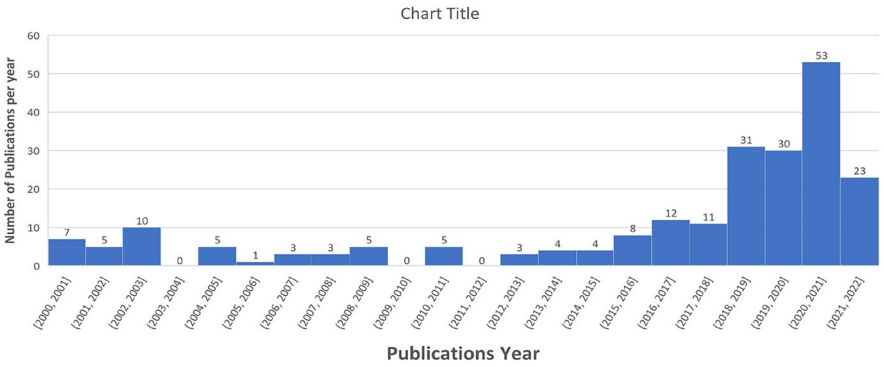

Fig. 3 Annual frequency of XAI and Law-themed academic publications

analysing potential research works in order to standardise reporting of Systematic Literature Reviews (SLR). However, methodological elaboration necessarily begins by detailing the search protocol procedure followed. In order to identify potential articles, the authors initially conducted a free text search on Google Scholar, capturing relevant search keywords for the comparatively comprehensive searches subsequently performed on article databases. The literature covered the period 2000 until 2022, which (as shown in Fig. 3) accounts for a period of singular development in AI and Law research. The literature search was conducted using the following keywords: "explainable artificial intelligence", "(XAI)", "XAI in Law", "AI in Law", and "Legal Reasoning". Secondly, a scoping study of all of the literature databases was effected, which resulted in the identification of a total of five relevant electronic databases: specifically, Scopus, IEEEXplore, Springer, the Association for Computing Machinery (ACM) digital library, and Jstor, all of which yielded relevant publications. Finally, the authors conducted a literature search on each individual database using the search protocol, supra. Inclusion and exclusion criteria were applied during the literature search. The inclusion criteria were published between 2001 and 2022, peer-reviewed articles, discussing XAI methodologies in law or evaluating methods, and challenges relating to XAI in Law, published as conference papers or in academic journals. The criteria for exclusion from the literature search were XAI articles that were not related to law, or to XAI domains, and articles that were not published in any peer-reviewed conference proceedings or journals. The steps involved in this process were identification, screening, eligibility assessment, and inclusion. The flow diagram of the above process is illustrated in Fig. 4, adapted from "Preferred Reporting Items for Systematic Reviews and Meta-Analyses" (PRISMA) diagram (Moher et al., 2009).

The relevant articles having been identified, these were combined, and duplicates were removed using Mendeley Reference Manager (Elston, 2019). The authors finally identified 137 articles which were selected as eligible for review. All searches were carried out in spring 2022. The subsequent review and discussion is detailed within the next section. However, prior to discussion it is worth noting, per Fig. 1,

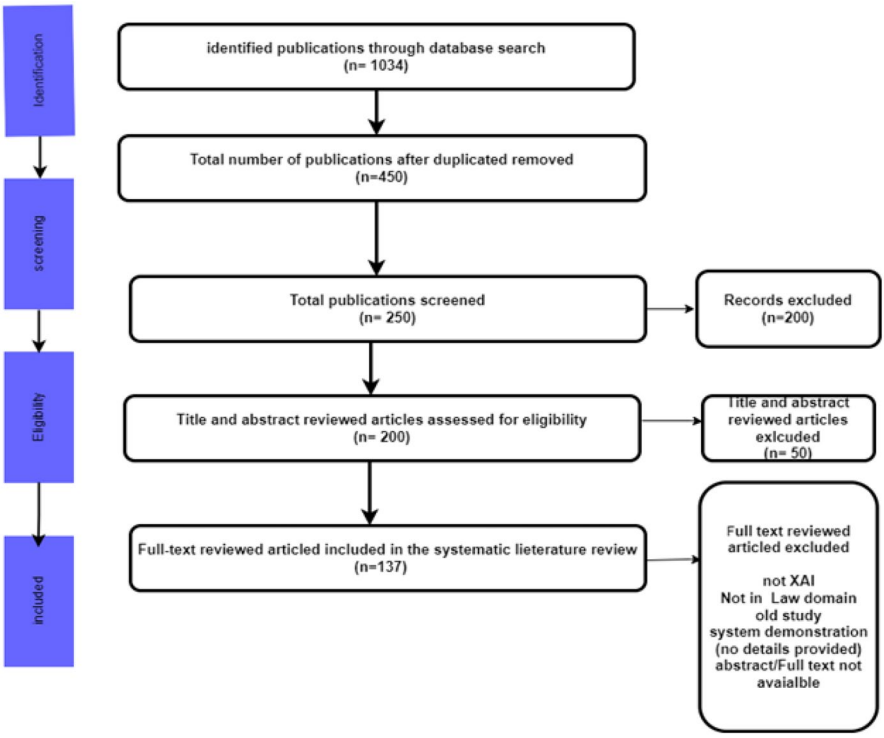

Fig. 4 Flow diagram of the research article selection process, adapted from Moher et al. (2009)

that research on the topic of XAI and law does not display an even chronological distribution. Rather, the topic remained relatively overlooked until 2018, when the graph shows a marked increase. This is reflective of more general trends in AI research and it is notable that the increasing focus on AI explainability within the legal field lags behind that of AI potential — and AI functionality within the legal field — and may be suggestive of an emerging understanding of the need to address opacity as a function of the shift to machine-learning technologies.

4 XAI and Law: Explanatory Categories

As stated, supra, this section elaborates upon the findings of the literature survey, commencing with a brief discussion of the key debates and areas of discursive uncertainty, before moving to a review of the divergent explanatory categories encountered within the surveyed works. During the course of the survey a series of technically grounded challenges were identified. Whilst it is necessary to treat of these key discursive areas, it should be reiterated that the focus of the current survey is upon the emergent interactions of explainable AI and heterogeneous legal sub-domains. Discussion of technical challenges related to data size, AI training, algorithmic fairness, and stakeholder evaluation criteria are necessarily included, albeit that they

sit on the periphery of the instant study. Firstly, it was noted that scalar challenges have emerged with the advent of legal "big data". These relate to volume, variety, velocity, and veracity (Antoniou et al., 2018). Such challenges complicate attempts to ground AI in the legal domain, leading to practical difficulties relating to (a) the handling of large data volumes (Antoniou et al., 2022; Devins et al., 2017); (b) the combination of streaming data with existing legal knowledge; (c) the integration of data from different sources and different formats; (d) the determination of provenance and qualitative improvement of available data; and (e) the effective combination of legal reasoning and analytic techniques. Further areas of concern were noted in relation to the embedding of legal concepts and language during AI development. Whilst domain-specific word embeddings and legal transformer models (LexGLUE) (Chalkidis et al., 2021) have routinely been employed, having been pre-trained using a large corpus of cases from the European Court of Human Rights (EctHR), Court of Justice of the European Union (CJEU), and Supreme Court of the Unite States (SCOTUS), it is noted that a number of similar models do not adequately capture the essential semantic content of legal text rendered in domain-specific documents (e.g. legislation and case law) since they were trained on generic corpora, e.g. Wikipedia articles, news stories, or randomly scanned web pages. Further challenges were noted in relation to algorithmic fairness, bias, and discrimination (Mehrabi et al., 2021). As has been exhaustively discussed in the extant literature, unfairness, bias, and discrimination metastasize across the body of XAI research, constituting a major challenge to the use of algorithms and automated decision-making systems across legal domains. Furthermore, with regard to end-user requirements, there is a lack of agreement over the necessary contours of stakeholder desiderata. Each stakeholder group privileges individual desiderata, including usability, understandability, trust, verification, fairness, morality, and accountability. However, the central aim of legal XAI is limited to satisfying the desiderata of legal stakeholders (Langer et al., 2021). Thus, having discussed the key debates, the focus of the survey discussion converges upon the central emergent plane of explainability — specifically the ability to render an explanation — and legal logic.

Explanations are an essential and inherent feature of the legal domain (Ashley & Rissland, 2003). Not only is argumentative reason-based discussion inherent in legal reasoning, but all parties have the right to an explanation when a decision is rendered in a court of law (Greenstein, 2022). In brief, proper justice requires decisions based on sound rationales. Symbolic AI and logic-based methods (intrinsically interpretable XAI methods) have been applied to model legal reasoning in the law domain and are well-suited to such applications, as they are highly interpretable by nature. However, differing methods have been used for modelling in the literature and these are categorised into four different types: rule-based legal reasoning, case-based legal reasoning, argumentation-based reasoning, and evidential (for the purposes of explainability "evidential" legal reasoning relates to the fundamental inferential processes of legal reasoning). There is also a fifth category composed of hybrid approaches.

Having reviewed the collected articles utilising the survey methodology described, this section provides a novel categorisation and discussion of the literature relative to the modes of legal reasoning employed, as shown in Fig. 5. Thus,

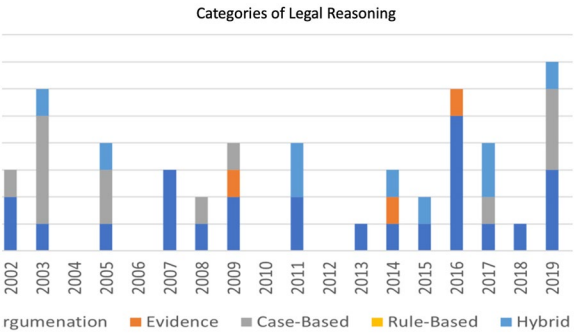

Fig. 5 Categories of legal reasoning approaches encountered across XAI and law publications (2000–2022)

the focus shifts to look at different categories of XAI methods used for modelling legal reasoning, and expounds upon the legal explainability — as opposed to causal explainability — of AI models. There are proliferating potentialities for the implementation of explainable AI in the legal domain. In a legal setting, this may be reactive and litigation-led, such as creating a decision tree that accurately reconstructs the decisions of a self-driving car's opaque algorithms in a product liability case, or the ways in which a probabilistic genotyping algorithm decides whether a suspect's DNA is included in a mixed DNA sample. These systems attempt to approximate the predictions made by an underlying model, whilst being interpretable (Richmond, 2021). Comparatively prosaic uses may involve the deployment of machine learning algorithms in administrative legal agencies, tasked with parsing and summarising voluminous bodies of documentation. Finally, proactive uses of AI may allow justice systems to implement machine learning to help them conduct courtroom adjudications and administrative decision-making. Such cognate iterations in the administrative law setting may also require judges to review an agency's use of machine learning algorithms. Finally, in the criminal justice setting, judges themselves may be the ones using those algorithms in order to predict recidivism. Whatever the features of

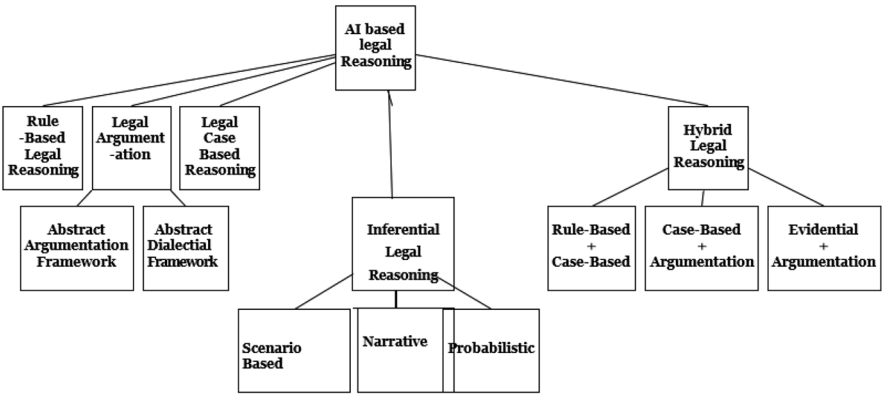

Fig. 6 Classification of modes of legal reasoning

the legal landscape, it is once more reiterated that an appreciation of the nuanced nature of legal procedure, its connection to multifarious forms of legal reasoning (as depicted in Fig. 6), and the potentials of different forms of AI application, all drawn from the survey of selected papers, may facilitate the future selection and development of specific forms of explainable AI. Discussion begins with the first category, rule-based legal reasoning.

4.1 Rule-Based Legal Reasoning

Legal reasoning describes the fundamental process by which a legal expert makes legal judgements using rules. Legal reasoning is a rule-guided activity, and allows for the application of legal rules to case interpretations. This type of reasoning is known as rule-based reasoning, and it can also be carried out by rule-based expert systems. The reasoning process is based on a series of if–then rule statements that are used to explain particular patterns in the given domain, such as legal norms (El Ghosh et al., 2017).

Rule-based systems remain the most prevalent legal AI systems. They model deductive reasoning, applying a rule of law to a given problem in order to obtain an answer A. The system declares A, based on the principle of law articulated by the legal authority that mandates it. The process of determining which rules should be applied — and how these should be interpreted — lies at the heart of legal reasoning (Mowbray et al., 2023). Such rule-based systems consist of three essential components: a set of rules (rule base), a fact base (knowledge base), and an interpreter for the rules (inference engine). Rules that reflect the content of knowledge-based sources are thereby applied, and matched with a set of facts, to deduce the conclusion using an inference engine.

Several rule-based legal systems have been discussed in the literature, whose explanations conform to the standard expert system form of "how, why, and what-if" explanations (Dattachaudhuri et al., 2021; Di Porto & Zuppetta, 2021; Kliegr et al., 2021). Such approaches were pioneered by MYCIN (Buchanan & Shortliffe, 1984), which reasoned by performing a form of logical inference on human readable symbols, and was able to provide a trace of its inference steps. More recently, authors proposed a theoretical framework of legal rule-based system for the criminal domain, named CORBS (El Ghosh et al., 2017). The system is founded on a homogeneous integration of a criminal domain ontology with a set of logic rules (Liu et al., 2021). Thus, CORBS stands as a unified framework that supports efficient legal reasoning. The framework relies upon a novel inference engine — the Semantic Rule Index — which can identify candidate rules alongside corresponding semantic rules, if any, and an inference controller that is able to guide the executions of queries and reasoning. Similarly, in Islam and Governatori (2018), the authors presented a rule-based (pre- and post-) reporting system (RuleRS) architecture to integrate databases — in particular relational databases (databases structured to recognise relations between stored items of information) — with a logic-based reasoner and rule engine, to assist in decision-making or create reports according to legal norms. The proposed system was demonstrated in a case study based on an online

child care management System, ChildSafeOMS, to automate the early identification of children at risk. The authors claim that the resulting RuleRS provides an efficient and flexible solution to the problem at hand by using defeasible inference. Thus, to summarise, it has been shown that AI-instantiated forms of rule-based legal reasoning exhibit the highest levels of explainability and interpretability across the surveyed literature, consonant with the most sensitive forms of legal decision-making. However, it should be noted that not all legal reasoning is predicated upon explicit rules. Much legal reasoning involves argumentation in relation to normative values. It is to this form of AI decision-making that discussion now turns.

4.2 Argument-Based Legal Reasoning

Argumentation is one of the most common legal reasoning methods in the law domain. It is based around the construction of arguments and counter-arguments (or "defeaters"), followed by the selection of the most acceptable of these (Amgoud & Cayrol, 2002b; Besnard & Hunter, 2001). Argumentation, as opposed to deduction, is an appropriate mode for reasoning with inconsistent knowledge, based upon the construction and the comparison of arguments. It can allow for reasoning in the face of uncertainty, and identify solutions when confronted with conflicting information. In particular, it should be possible to use this approach to assess the reason why a putative fact resonates, in the form of argument, and to combine these arguments to evaluate the level of certainty (Možina et al., 2007). Argumentation has strong explanatory capabilities, as it can translate the decision of an AI system to an argumentation process, which shows step-by-step how the system concludes the result (Vassiliades et al., 2021). Turning to the underlying argumentation theory, the first argumentation framework was proposed by Toulmin (1958). It was designed to address the structure of commonplace arguments, and was followed by several other forms of argumentation framework (Šešelja & Straßer, 2013; Labrie & Schulz, 2014; Charwat et al., 2015; Schulz & Toni, 2016; Kökciyan et al., 2017; Liepiņa et al., 2020; Zhong et al., 2019; Neil et al., 2019; Lamy et al., 2019; Yu & Chen, 2023). In recent years, the abstract argumentation framework proposed by Dung (1995) has become the benchmark for the application of formal argumentation models to legal argumentation, of the sort encountered in decisions of the US Supreme Court (Dung, 1995). The idea behind Dung's work is that given a set of arguments, where some arguments attack others, in order to determine whether an argument can ultimately be accepted or not, it is not sufficient to look at the interim stages of an argument, because these could be defeated by other arguments (Burgemeestre et al., 2011; Dunne et al., 2011). The system must aim to find the arguments that can ultimately be accepted. In Prakken et al. (2015), the authors provide a legal formalisation of argumentation theory: the legal case is first fed into the ASPIC framework that tries to produce defeasible rules, which are collated as arguments. The authors then provided (Amgoud & Cayrol, 2002b) a revised proof theory in terms of AND/OR trees, verifying whether a given argument is acceptable, and conforms with the dialectical form of argumentation. On that note, it should be highlighted that several types of argumentation approaches have been proposed in the literature, based upon

different perspectives on arguments and semantics (Caroprese et al., 2022; Vassiliades et al., 2021). Many of these are derived from the basic framework proposed by Dung (1995), whose main concepts are the argumentative proposition, its subsequent extension, and the particular category of extension. The taxonomy of approaches includes the Bipolar Argumentation Framework (BAF), the Label-Based Argumentation Framework (LBAF), Structured Argumentation Framework (SAF), the Quantitative Bipolar Argumentation Framework (QBAF), the Probabilistic Bipolar Argumentation Framework (PBAF), and Weighted Argumentation Frameworks (WAFs). Not all of these extension argumentation approaches have, as yet, been applied in the law domain. However, recently, a powerful generalisation of Dung's abstract argumentation framework, Abstract Dialectical Framework (ADF), has been developed and successfully employed in the legal field (Al-Abdulkarim et al., 2014, 2016), demonstrating how these structures provide an excellent framework for reasoning with legal cases. Similarly, in Collenette et al. (2020), the authors presented an argumentation-based representation of Article 6 of the European Convention on Human Rights, which is the right to a fair trial. The representation is an Abstract Dialectical Framework produced using the ANGELIC methodology, whereby the domain is represented as a tree, with the root, or parent, node being a verdict, followed by children, which represent the issues, whose children themselves are abstract factors. This framework was written in Prolog to allow for domain-specific reasoning. The Prolog program was tested using cases involving Article 6 (and further additional articles) that were decided in the European Court of Human Rights (ECtHR).

4.3 Legal Case-Based Reasoning Methods

Both rule-based and argument-based forms of reasoning have proved relatively easy to translate to the legal field, given the adherence to argumentation and dialectics in many sub-domains. Another significant strand of research within the surveyed literature, however, applies a problem-solving approach — not referable to strict rules — known as Legal Case-Based Reasoning (LCBR) (Keane & Kenny, 2019; Heras et al., 2009). Works in this category conform to the analogical reasoning of the courtroom in which a judge reasons with instant and prior cases (drawn from a case base), finding similarities and differences between them (Rissland et al., 2005). In legal terms, this is based on the *stare decisis* concept, which mandates that similar cases should be decided similarly. In LCBR a collection of domain-dependent, legally significant features are defined, and a prior case is judged to be relevant (and possibly binding) depending on the degree of match between the prior case's features and those of the present case. Legal case-based reasoning has been formalised for the purposes of computer reasoning as a four-step process. In general applications it is iterated as follows:

1. Retrieve: given a target problem, retrieve relevant cases from memory. A given case thus consists of a problem, its solution, and (typically) annotations about how the solution was derived.

2. Revise: having mapped the previous solution to the target situation, test the new solution in the real world (or a simulation) and, if necessary, revise.
3. Retain: after the solution has been successfully adapted to the target problem, store the resulting experience as a new case in memory, and reuse where necessary.

The first legal CBR model that addressed all of these features was HYPO (Ashley, 1989), developed for analysing cases, and constructing legal arguments, in the domain of trade secrets law. In HYPO, the set of shared factors between two cases are called "relevant similarity", whilst the set of unshared factors are called "relevant differences". Unshared factors can be used for pointing out the two cases should be decided differently. CATO was intended to address a legal task central to legal training and legal reasoning, specifically to support law students in their attempts to distinguish cases. After the introduction of HYPO and CATO, a variety of approaches have evolved, such as IBP (Aleven, 1997; Bourcier, 2003), which determines the underlying issues of a case and, on the basis of these, predicts the verdict of case-based legal processes. IBP was followed by GREBE (Generator of Recursive Exemplar-Based Explanations) (Branting et al., 2021), CABARET (Rissland & Skalak, 1991), and BankXX (Rissland et al., 1996). These newer LCBR methods have brought modifications and advancements in LCBR modelling. For example, the authors in Zheng et al. (2021) proposed a logical comparison approach, which logically generalised the formulas involved in case comparison, and their approach to identifying analogies, distinctions, and relevances. This approach is extended to HYPO-style comparisons (where distinctions and relevance are not separately characterised) and to the temporal dynamics of case-based reasoning modelling real-world cases. Notably, the authors claimed that such case-based model formalism is capable of refining the comparisons inherent to case-based reasoning. Further, in Branting (2003) the authors described a reduction-graph model of legal precedent that makes explicit the theory of decision of the precedent. This model accounts for the phenomenon that the theory under which a precedent is decided determines its relevance to subsequent cases. The authors demonstrate this reduction graph model of precedent using GREBE legal analysis architecture, which uses a relational representation of facts, structure-matching for similarity assessment, and a control strategy that treats rules, precedents, and semantic relations in a uniform fashion. Further, legal case-based reasoning approaches using ontologies have been developed. These facilitate the exchange and re-use of knowledge and information amongst knowledge bases, make the assumptions about concepts explicit so that the algorithmic program can reason with them, and manage relations and distinctions amongst concept types, helping to generate natural language explanations. For example, the authors in Wyner (2008) developed an ontology-based LCBR theoretical framework by using Web Ontology Language (OWL), where the model reasons from cases that act as legal precedents (PCs) to argue for a decision for a plaintiff (P) or defendant (D) in a current case (CC). A variety of extended techniques have been used to formalise further factors in legal case-based reasoning, for instance, dialectical arguments (Bench-Capon et al., 2000), context-related frameworks (Hafner & Berman, 2002), ontologies in OWL (Wyner, 2008), the ASPIC + framework (Modgil &

Prakken, 2014), reasoning models (Horty & Bench-Capon, 2012), abstract argumentation (Prakken, 2010), abstract dialectical frameworks (Al-Abdulkarim et al., 2016), and abductive logical programming. It is notable from the above survey that CBR methods can be used in situations where the legal rules and the evidential matrix are less explicitly rendered, such as in cases involving administrative decisions based solely upon a credibility assessment. Thus, case-based methods may be suited to legal fields in which the evidence is judged holistically, as opposed to atomistic ally, and is consequently limited in terms of its structure and interpretability, This will be discussed, infra. However, the survey now turns to a further category of methods based upon legal inference, which for ease of description fall under the title of evidential approaches.

4.4 Inferential Legal Reasoning Methods

Evidential legal reasoning based upon rational inference is one of the most salient features of the legal domain. The aim is to reason with help of evidential data, by which is meant the primary sources of evidence the existence of which cannot be sensibly denied, e.g. witness statements made in court, forensic expert reports presented to the trier-of-fact. Traditionally, there are three main approaches to modelling reasoning with evidence: narrative/scenario based, probabilistic, and argumentative approaches (Verheij, 2017). A narrative approach to reasoning with legal evidence revolves around the concept of "scenarios". Different hypothetical scenarios about what has happened are considered side-by-side, and considered in light of the evidence. Scenario analysis enables the coherent interpretation of all evidence, whilst the second category — the probabilistic approach — is suitable for analysing statistical evidence, and in the argumentative approach, a structured constellation of evidence, reasons, and hypotheses are considered. Usually, the evidence that gives rise to reasons for and against the possible conclusions is considered (Verheij, 2017). A forensic expert might use probability to report his findings whilst a judge or jury might be more likely to utilise argumentative or narratives. Bayesian networks are the most commonly used for probabilistic approaches (Balding, 2011). It has become the most popular probabilistic tool for working with forensic evidence. Finally, the argumentative approach focusses upon a matrix of legal inferences connecting the evidence to the ultimate question at issue. In practice, our survey shows that elements from each approach are often combined. In Vlek et al. (2014) the authors thus proposed a framework combining argumentative, narrative, and probabilistic techniques. The authors argue that integrating these three approaches could arguably enhance the communication between an expert and a judge or jury by modelling the evidence and a number of relevant scenarios in a Bayesian network. This design method is evaluated by means of an extensive case study concerning the notorious Dutch case of the Anjum murders. Similarly, in Vlek et al. (2016) the authors proposed an evidential reasoning approach by combining probabilistic and narrative techniques that allows a judge or jury to work with statistical information whilst considering the whole case in terms of narrative techniques. Recently, the authors in Biedermann et al. (2020) have shown how statistical decision theory

can be fruitfully applied as an analytical and a normative tool to the decision problem that forensic experts routinely face. Decision theory is an extensively discussed topic in legal literature, but its use in forensic science is more recent. The authors thus compare different ways of stating forensic identification decisions in decision-theoretic terms and explain their conclusions. However, it is in the field of evidential legal reasoning that the issue of "legal logics" is thrown into greatest relief for, as will be demonstrated below, differing logics are at play in different legal sub-fields, and these differing forms, with their varied potentials and affordances, are suited to particular forms of XAI, with differing degrees of explainability.

As stated in the introduction, whilst the literature on legal XAI has proliferated in recent years (see Fig. 3), existing scholarship also spans a range of legal sub-domains and modality boundaries (see Fig. 5). From a legal perspective this raises the question of generalizability of existing XAI algorithms across different legal domains and legal systems. The mapping and analysis in the previous sections suggest a need to move beyond bounded definitions of explainability, and to embrace explanatory pluralism as a concept for mapping diverse legal uses of XAI. Yet, just as attention to different algorithmic designs matters in terms of legal explainability, attention to differences within the legal domain is essential for how well a given technical approach is likely to work in a given area of law. In this section, the authors thus provide a closer examination of how the above categories map onto underlying dimensions related to legal systems, domain areas, evidentiary standards, modes of adjudication and decision-making, and legal reasoning. This opens up a structured way for scholars working in this area to evaluate whether existing algorithms can usefully be applied in other domains and jurisdictions. As in the previous section, this serves as the basis for a more coherent and foundationally robust approach to law and XAI. At the most basic level, the disciplinary outlook seems to impact the ontological (what exists for people to know about) and epistemological premises from which analysis is advanced. In the instant survey, papers from the data science field tend to apply a realist ontological perspective, whilst a significant proportion of papers written by scholars in law and the humanities view legal and scientific endeavours as more socially constructed (Rotolo & Sartor, 2023). A similar divide emerges in relation to epistemology. In overall form, the surveyed works drawn from the data science field exhibited a scientific positivist approach to a not-insignificant degree. In other words, following the tradition of Reichenbach and Popper, their approach to the study of society relies solely on empirical scientific evidence, such as controlled experiments and statistics, and they do not stray beyond these boundaries. Nonetheless, a significant number of papers from the data science field advanced explanations that operate on the social — as opposed to the purely technical — level (Rosengrün, 2022). Others proffered mechanistic and technical explanations (Di Porto & Zuppetta, 2021), whilst a significant number proffered explanations based upon regulatory requirements (Council of the European Union, 2021; Gutierrez et al., 2023). Such disciplinary commitments and divergencies should be appreciated when interdisciplinary literature. Having dealt with the most typical legal applications and their onto-epistemic dimensions, discussion now turns to hybrid forms.

4.5 Hybrid Legal Reasoning

Discussion now turns to ancillary categories of legal reasoning, or what might be referred to as hybrid approaches (Walton, 2019). The integration of two or more different knowledge representation methods is an oft-used research strategy (Rissland & Skalak, 1989), the underlying assumption being that complex problems can be easier solved with hybrid systems (Hamdani et al., 2021; Marques Martins, 2020). One of the most popular types of integration within the legal domain involves the combination of rule-based with case-based reasoning approaches. The rule-based approach has the advantage of structuring the explanation according to the underlying statute or legal doctrine, but tends to be rather prescriptive and requires considerable knowledge engineering effort in constructing the rule base (bearing in mind that rules represent general knowledge of the domain, whereas cases represent specific knowledge). Rule-based systems solve problems from scratch, whilst case-based systems use pre-stored situations to deal with similar new instances. Therefore, the integration of both approaches is natural and often useful. One of the first real hybrid CBR–RBR (case-based and rule-based reasoning) systems is yet another descendant of HYPO, the CAse-BAsed REasoning Tool (CABARET) system (Rissland & Skalak, 1991), which produces legal arguments in the domain of American tax law. CABARET has a domain-independent architecture that includes two reasoners (one case-based and the other rule-based) that are managed by an agenda controller, which uses heuristic rules to dynamically alternate the control over them. Yet another important hybrid CBR–RBR system is the GeneratoR of Exemplar-Based Explanations (GREBE) (Branting & Branting, 2000), which pioneered using the justifications of legal cases to create new arguments. GREBE is a legal analysis system that reasons with portions of precedent cases in the domain of the State of Texas worker's compensation law. Like CABARET, GREBE is a hybrid CBR/RBR program that reasons with both rules and cases. For instance, it can create case analogies when the rules run out, or otherwise fail to show that a legal term has been satisfied. GREBE used a heuristic measure of argument strength to rank the arguments for and against a given conclusion. Its designer, Branting, also tackled the issue of how to evaluate a CBR system satisfactorily (Branting & Branting, 2000). Thus, GREBE used a heuristic measure of argument strength to rank the arguments for and against a given conclusion.

Moving from rule-based CBR hybrids to argument-based CBR hybrids it was noted that several studies used LCBR and the argumentation framework together for legal reasoning. This allows for a more robust ontology to be developed within the CBR paradigm. In Wyner and Bench-Capon (2007), a number of novel legal case-based argumentation schemes are specified, such as Aleven and Ashley's CATO (Aleven, 1997), where the model attempts to determine how and in what way a precedent case does (or does not) argue in support of a determination in the current case. Meanwhile, Al-Abdulkarim et al. (2016) explored the use of Abstract Dialectal Argumentation Frameworks to formalise factors and showed that they can provide a natural way to express formal reasoning with legal cases using factors developed through important practical systems such as the afore-mentioned HYPO (Ashley, 1989), CATO (Aleven, 1997), and IBP (Bourcier, 2003). Similarly,

Wyner et al. (2011) reasoning with legal cases in terms of argumentation schemes was again proposed. Factors were formalised using the ASPIC + framework (Modgil & Prakken, 2014) and begin by modelling the style of reasoning developed in the CATO project (which describes cases using factors) and then extends the account to accommodate the dimensions used in the HYPO project. Yet another novel hybrid model of legal reasoning was proposed in Bex and Verheij (2011), by expanding the argumentative approach to evidential reasoning in order to encompass the entire reasoning process in a case, from evidence through facts to legal implications. It is an extension of their proposed hybrid theory of reasoning with evidence. The authors in Bex and Verheij (2011) argue that the process from evidence to established facts and from established facts to legal implications cannot be isolated from the factual component of legal reasoning. Thus, it has been demonstrated that the categories of AI-powered legal reasoning are not discrete but offer the potential for the development of hybrid forms that fit with the different facets of legal reasoning. Finally, our discussion turns to post hoc explanatory models developed to offer explanations in respect of the most opaque forms of AI.

4.6 Post hoc XAI Algorithms

As stated, supra, many state-of-the-art AI models are opaque systems that reason without transparency. To the extent that their decision-making is not amenable to explanation, these are inherently unsuitable in the context of AI and law. Neural networks, for example, are known to perform extremely well but nonetheless behave opaquely. Hence, explanation techniques have been developed to "open the black box". The central XAI methods which have been developed, in order to interpret the black box model decisions, can be divided into three categories, namely Intrinsically Interpretable approaches, Post hoc approaches, and example-based approaches and XAI methods such as LIME (Ribeiro et al., 2016), SIDU (Muddamsetty et al., 2022), and SHAP (Lundberg & Lee, 2017), which allow users to look "inside" the black box by demonstrating which parts of the input are important in the system's decision-making process. SHAP (SHapley Additive exPlanation) is an explainable AI framework that explains the output of a machine learning system based on the idea of Shapley values from game theory. LIME creates explanations by perturbing individual instances and using these to learn interpretable sparse linear models that approximate the system's decision-making. These state-of-the-art XAI methods are applied in several sensitive domains, such as medicine (Zhang et al., 2022) and finance (Kuiper et al., 2022). However, the use of existing XAI algorithms is comparatively limited in the law domain, though XAI methods have been utilised in legal practice for the purpose of preliminary evaluation (Steging et al., 2021). In a recent study, lawyers were tasked with assessing both LIME and SHAP in a routine legal text classification task, grading both explanation methods similarly and recommending the use of systems with greater explainability to assist their work (Górski et al., 2021). The most recent post hoc explainable AI techniques (Lundberg & Lee, 2017; Ribeiro et al., 2016) have developed in order to identify the dataset features that had the greatest impact on the classification process. Another recent

paper (Górski et al., 2021) uses a popular image processing technique, Grad-CAM (Selvaraju et al., 2017), to demonstrate explainability in relation to legal texts. Similarly, in Dadgostari et al. (2021) the authors tested an approach to exploiting the scalability of machine learning whilst attempting to retain explanatory capability. The evaluation of the approach, Attention Network-based Predictions (ANP), failed to establish significant improvement in decision accuracy but offered positive indications relating to decision support in the form of highlighting case text based on attention weights. In summary, the experiments do not conclusively establish that this approach is not workable, but do suggest the limits of explanation and justification based purely on semantic codings without explicit reference to legally relevant concepts. The central post hoc approaches are tabulated in order to aid further research, showing (in Table 1) the particular XAI algorithm, method, purpose, and the dataset utilised.

In summary, it can be stated that rule-based systems comprise the most prevalent body of legal AI expert systems. These benchmark approaches model deductive reasoning, applying a rule of law to a given problem in order to obtain an answer. Such approaches are to be recommended in tightly delineated legal fields with set legal rules and procedural forms. In contrast, argument-based approaches were found to comprise a promising suite of approaches for reasoning with inconsistent knowledge, based on the construction and the comparison of arguments. These methods readily allow for reasoning in the face of uncertainty, and identification of solutions when confronted with conflicting information. A wholly different strand of research was embodied by Case-Based Reasoning methods. These were found to fit comfortably with the legal doctrine of stare decisis, which mandates that similar cases should be decided similarly. Thus, in LCBR a collection of domain-dependent, legally significant features are defined, and a prior case is judged to be relevant (and possibly binding) depending on the degree of match between the prior case's features and those of the present case, enabling researchers to solve problems and make predictions in circumstances where a large case-base exists, turning to "evidential" legal reasoning methods, central to legal inference, and comprising three sub-categories: narrative or scenario based analysis, probabilistic approaches, and argument-based approaches. The survey revealed that each of these sub-categories possesses strengths and weaknesses which make their application more, or less, suited to particular legal sub-fields: for example, probabilistic approaches are most suited to forensic exposition. Discussion of hybrid forms offered added potentials, given that the combination of rule-based, argument-based, and case-based approaches could parse with the different aspects of legal reasoning, as applied to both legal norms and rules, and fact-based inferences grounded in general knowledge and expertise. Finally, it was demonstrated that, in respect of the most opaque forms of AI such as neural nets, explainability remains possible, though this remains tightly circumscribed, generating only technical explanations which do not yet fulfil the requirements of the legal domain. Once again, it is clear that differing forms of AI may offer higher accuracy but at the expense of explainability, the latter being a crucial factor in legal reasoning and adjudication (or administrative decision-making). To aid future research, the foregoing discussion has been tabulated, in Table 2, showing all of the works referenced in Section 4, the central mode of legal reasoning relative

Table 1 Post hoc XAI algorithms applied within legal domain

Article	Year	XAI algorithm	XAI method	Purpose	Data type
Branting et al. (2021)	2020	Attention Network-based Prediction (ANP)	Post hoc	To demonstrate significance and range of explainability regarding use of ML models in contract/tort litigation	World Intellectual Property Organization (WIPO) domain-name dispute decisions
Steging et al. (2021)	2021	LIME, SHAP	Post hoc	To investigate whether explainable AI techniques can be used to discover the unsound rationale as it is actually used by the trained system	Welfare Benefit domain dataset
Górski et al. (2021)	2021	GRAD-CAM, LIME, SHAP	Post hoc	To assess the performance of different explainability methods in explaining the predictions for a legal text classification problem	Statutory Interpretation –Identifying Particular (SIIP) datasets
Górski et al. (2021)	2021	GRAD-CAM	Post hoc	To showcase the explainability concept for legal texts	Post-Traumatic Stress Disorder (PTSD) and Statutory Interpretation-Identifying Particular (SIIP) datasets

Table 2 Summary of XAI literature

Publication(s)	Year	Mode of legal reasoning	Publication type	Legal sub-domain	Evaluative outputs
Bench-Capon et al. (2000)	2000	Argumentation	Technical	Dialectical argument	Manual assessment
Besnard and Hunter (2001)	2001	Argumentation	Theoretical framework	Deductive argument	Comparative analysis of argument systems
Bench-Capon and Sartor (2001)	2001	Hybrid (argumentation/LCBR)	Theoretical framework	Precedent/case theory	Theoretical evaluation
Amgoud and Cayrol (2002a)	2002	Argumentation	Theoretical framework	Dialectical argument	Evaluation absent
Amgoud and Cayrol (2002b)	2002	Argumentation	Theoretical framework	Arguments and counter-arguments	Evaluation absent
Ashley (2002)	2002	Case-based reasoning	Theoretical framework	Case-based legal argument	Evaluation absent
Verheij (2003)	2003	Argumentation	Theoretical framework	Dialectical argument	Evaluation absent
Aleven (2003)	2003	Case-based reasoning	Technical	Basic argumentation skills	Predictive accuracy and effectiveness
Branting (2003)	2003	Case-based reasoning	Technical	Civil litigation	Manual assessment
Bench-Capon and Sartor (2003)	2003	Case-based reasoning	Theoretical framework	Liability for animals	Theoretical evaluation
Ashley and Rissland (2003)	2003	Case-based reasoning	Theoretical framework	Civil litigation	Evaluation absent
Rissland et al. (2005)	2005	Case-based reasoning	Survey paper	General legal reasoning	Evaluation absent
Možina et al. (2007)	2007	Argumentation	Technical	Social welfare	Statistical accuracy
Atkinson and Bench-Capon (2007)	2007	Argumentation	Technical	General legal reasoning	Critical evaluation
Amgoud et al. (2008)	2008	Argumentation	Survey paper	General legal reasoning	Evaluation absent
Wyner (2008)	2008	Case-based reasoning	Theoretical framework	General legal reasoning	Evaluation absent
Amgoud and Prade (2009)	2009	Argumentation	Theoretical framework	Criminal law	Evaluation absent
Atkinson and Bench-Capon (2019)	2009	Logical inference	Theoretical framework	General legal reasoning	Evaluation absent
Heras et al. (2009)	2009	Case-based reasoning	Theoretical framework	Criminal law	Evaluation absent
Dunne et al. (2011)	2011	Argumentation	Theoretical framework	General legal reasoning	Evaluation absent
Burgemeestre et al. (2011)	2011	Argumentation	Theoretical framework	Regulatory compliance	Evaluation absent

Table 2 (continued)

Publication(s)	Year	Mode of legal reasoning	Publication type	Legal sub-domain	Evaluative outputs
Bex and Verheij (2011)	2011	Hybrid (inference/argumentation)	Theoretical framework	Criminal law	Evaluation absent
Wyner et al. (2011)	2011	Hybrid (argumentation/CBR)	Theoretical framework	Liability for animals	Evaluation absent
Šešelja and Straßer (2013)	2013	Argumentation	Theoretical framework	Scientific efficacy and regulation	Evaluation absent
Labrie and Schulz (2014)	2014	Argumentation	Survey paper	General legal reasoning	Evaluation absent
Al-Abdulkarim et al. (2014)	2014	Hybrid (argumentation/CBR)	Theoretical framework	General legal reasoning	Comprehensive evaluation
Vlek et al. (2014)	2014	Logical inference	Technical	Criminal law	Case study evaluation
Charwat et al. (2015)	2015	Argumentation	Survey paper	General legal reasoning	Evaluation absent
Al-Abdulkarim et al. (2015)	2015	Hybrid (argumentation/CBR)	Theoretical framework	General legal reasoning	Evaluation absent
Verheij (2016)	2016	Argumentation	Theoretical framework	Tort	Evaluation absent
Schulz and Toni (2016)	2016	Argumentation	Theoretical framework	General legal reasoning	Evaluation absent
Bex and Walton (2016)	2016	Argumentation	Theoretical framework	General legal reasoning	Evaluation absent
Al-Abdulkarim et al. (2016)	2016	Argumentation	Technical	Taxation	Performance evaluation
Vlek et al. (2016)	2016	Logical inference	Technical	Criminal law	Case study evaluation
Kökciyan et al. (2017)	2017	Argumentation	Technical	Security/privacy	Case study evaluation
Bench-Capon and Atkinson (2017)	2017	Case-based reasoning	Theoretical framework	Citizenship	Evaluation absent
El Ghosh et al. (2017)	2017	Rule-based reasoning	Theoretical framework	Criminal law	Evaluation absent
Verheij (2017)	2017	Hybrid (logical inference/argumentation)	Theoretical framework	Forensic evidence	Evaluation absent
Branting (2017)	2017	Hybrid (logical inference/data-centric)	Theoretical framework	General legal reasoning	Evaluation absent
Islam and Governatori (2018)	2018	Rule-based reasoning	Technical	Food and drug regulation	Quantitative evaluation
Zhong et al. (2019)	2019	Argumentation	Technical	Criminal law	Expert appraisal
Neil et al. (2019)	2019	Argumentation	Theoretical framework	Criminal law	Evaluation absent

Table 2 (continued)

Publication(s)	Year	Mode of legal reasoning	Publication type	Legal sub-domain	Evaluative outputs
Lamy et al. (2019)	2019	Case-based reasoning	Technical	Medical litigation	Statistical evaluation
Walton (2019)	2019	Hybrid (logical inference/argumentation)	Theoretical framework	Criminal law	Evaluation absent
Keane and Kenny (2019)	2019	Case-based reasoning	Survey paper	General legal reasoning	Evaluation absent
Liepiņa et al. (2020)	2020	Argumentation	Theoretical framework	General legal reasoning	Evaluation absent
Collenette et al. (2020)	2020	Argumentation	Technical	Human rights	Critical evaluation
Marques Martins (2020)	2020	Hybrid (argumentation/logical inference)	Technical	Civil litigation	Statistical evaluation
Biedermann et al. (2020)	2020	Logical inference	Theoretical framework	Forensic evidence	Evaluation absent
Vassiliades et al. (2021)	2021	Argumentation	Survey paper	General legal reasoning	Evaluation absent
Zheng et al. (2021)	2021	Case-based reasoning	Theoretical framework	General legal reasoning	Evaluation absent
Kliegr et al. (2021)	2021	Rule-based reasoning	Survey paper	General legal reasoning	Evaluation absent
Hamdani et al. (2021)	2021	Hybrid (rule-based reasoning/datacentric)	Technical	GDPR	Statistical evaluation
Dattachaudhuri et al. (2021)	2021	Rule-based reasoning	Technical	General legal reasoning	Statistical evaluation
Liu et al. (2021)	2021	Rule-based reasoning	Technical	General legal reasoning	Quantitative evaluation
Di Porto and Zuppetta (2021)	2021	Rule-based reasoning	Technical	GDPR	Evaluation absent
Caroprese et al. (2022)	2022	Argumentation	Survey paper	General legal reasoning	Evaluation absent
Yu and Chen (2023)	2022	Argumentation	Theoretical framework	Citizenship	Evaluation absent
Mowbray et al. (2023)	2022	Rule-based reasoning, XAI	Technical	NSW Department of Fair Trading *Community Gaming Regulation 2020*	Evaluation absent

to each, the publication type, legal sub-domain dealt with, and, where present, the mode of evaluation.

To conclude, in this study, the authors surveyed and categorised how — and in what diverse ways — explanatory techniques are used in legal practice, administration, and adjudication. The authors noted that the domain expertise on which AI designs are founded is not only implicit and opaque, but under-theorised within the extant literature. Such a lack of theorisation as regards the nature of legal reasoning may generate obstacles when implementing AI for automated decision-making tasks within the legal field. Thus, the level of opacity is linked to the degree to which AI-powered adjudication and decision-making tools must found upon a stock of domain expertise. The heterogeneous nature of legal reasoning is a salient factor which requires elaboration, and should be accounted for when designing AI-driven decision-making systems for the legal field, as the mode of legal reasoning will necessarily condition the explainability level of the output. It is thereby hoped that adjudicators, decision-makers, researchers, and practitioners can gain unique insights into explainability, select more appropriate types of AI, and utilise the survey as the basis for further research within the field.

5 Concluding Remarks

The unique methodological approach of this survey has been to avoid approaching the field of AI and Law from a monolithic perspective, viewing adjudication and decision-making as a "one-size-fits-all" approach which shows little divergence in its various sub-domain iterations (administrative, criminal, civil, etc.) excepting those conditioned by interposing procedural rules. Rather, this study has demonstrated, by reference to the literature, that particular divergent — though overlapping — forms of legal reasoning pertain to particular fields. Furthermore, the dominant forms of logic (abductive, deductive, inductive) map to particular modes of fact finding (atomistic, holistic, analogical, and hypothetico-deductive). Finally — and this is the core finding of the survey paper — that these divergent logical forms and modes of reasoning require different levels of explainability, and thereby determine the suitability of particular forms of AI model (deep-learning, case-based reasoning, etc.). Therefore, in term of future directions for AI and legal research, as approached from the legal perspective, it will be crucial to remain mindful of the interactions and relations between the dominant form of reasoning, mode of fact-finding, and level of explainability sought, with regard to the choice of AI model. Thus far, this area remains almost entirely unexplored.

Author Contribution Karen Richmond and Satya Muddamsetty contributed to the design and implementation of the survey, to the collection of data, to the analysis of the results, and to the writing of the manuscript. Henrik Palmer Olsen, Thomas Gammeltoft-Hansen, and Thomas Moeslund contributed to the writing of the manuscript.

Funding Open access funding provided by Royal Library, Copenhagen University Library. This work was conducted as a part of the XAIfair project funded by a Villum Synergy from the Villum Foundation. The total grant amount is DKK 2,999,526. The project period runs from 1 December 2021 to 30 November 2023.

Data Availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Informed Consent Not applicable.

Competing Interests The authors declare no competing interests.

References

Aamodt, A. (1991). *A knowledge-intensive, integrated approach to problem-solving and sustained learning*. Norway: Universitetet i Trondheim. Doctoral dissertation.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Al-Abdulkarim, L., et al. (2014). *Abstract dialectical frameworks for legal reasoning* (pp. 61–70). IOS Press.

Al-Abdulkarim, L., et al. (2016). A methodology for designing systems to reason with legal cases using abstract dialectical frameworks. *Artificial Intelligence and Law, 24*, 1–49.

Al-Abdulkarim, L., Atkinson, K., & Bench-Capon, T. (2015). Factors, issues and values: Revisiting reasoning with cases. *Proceedings of the 15th international conference on artificial intelligence and law* (pp. 3–12).

Aletras, N., et al. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science, 2*, e93.

Aleven, V. (2003). Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. *Artificial Intelligence, 150*(1–2), 183–237.

Aleven, V. A. (1997). *Teaching case-based argumentation through a model and examples*. Pittsburgh: University of Pittsburgh.

Alikhademi, K., et al. (2022). A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law,* 1–17.

Amgoud, L., & Cayrol, C. (2002a). Inferring from inconsistency in preference-based argumentation frameworks. *Journal of Automated Reasoning, 29*, 125–169.

Amgoud, L., & Cayrol, C. (2002b). A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence, 34*, 197–215.

Amgoud, L., et al. (2008). On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems, 23*(10), 1062–1093.

Amgoud, L., & Prade, H. (2009). Using arguments for making and explaining decisions. *Artificial Intelligence, 173*(3–4), 413–436.

Antoniou, G., et al. (2022). Explainable reasoning with legal big data: A layered framework. *IfCoLoG Journal of Logics and Their Applications, 9*(4), 1155–1170.

Antoniou, G., et al. (2018). Legal reasoning and big data: Opportunities and challenges. *Legal Reasoning and Big Data: Opportunities and Challenges*.

Ashley, K. D. (1989). Modelling legal argument: Reasoning with cases and hypotheticals.

Ashley, K. D. (2002). An AI model of case-based legal argument from a jurisprudential viewpoint. *Artificial Intelligence and Law, 10*(1–3), 163–218.

Ashley, K. D., & Rissland, E. L. (2003). Law, learning and representation. *Artificial Intelligence, 150*(1–2), 17–58.

Atkinson, K., & Bench-Capon, T. (2007). Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence, 171*(10–15), 855–874.

Atkinson, K., & Bench-Capon, T. (2019). Reasoning with legal cases: Analogy or rule application? *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (pp. 12–21). ABC.

Atkinson, K., et al. (2020). Explanation in AI and Law: Past, present and future. *Artificial Intelligence, 289*, 103387.

Balding, D. J. (2011). *Bayesian networks and probabilistic inference in forensic science*. Oxford University Press.

Barredo, P., Hernández-Orallo, J., Martınez-Plumed, F., & h Éigeartaigh, S. O. (2020). The scientometrics of AI benchmarks: Unveiling the underlying mechanics of AI research. *Evaluating progress in artificial intelligence (EPAI 2020).* ECAI.

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data,* 39.

Bench-Capon, T., & Sartor, G. (2001). Theory based explanation of case law domains: 38. *Proceedings of the 8th international conference on artificial intelligence and law* (pp. 12–21).

Bench-Capon, T., & Sartor, G. (2003). A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence, 150*(1–2), 97–143.

Bench-Capon, T. J., & Atkinson, K. (2017). Dimensions and values for legal CBR. *JURIX* (pp. 27–32).

Bench-Capon, T. J. M., et al. (2000). A method for the computational modelling of dialectical argument with dialogue games. *Artificial Intelligence and Law, 8*, 233–254.

Besnard, P., & Hunter, A. (2001). A logic-based theory of deductive arguments. *Artificial Intelligence, 128*(1–2), 203–235.

Bex, F., & Verheij, B. (2011). Legal shifts in the process of proof. *Proceedings of the 13th International Conference on Artificial Intelligence and Law*.

Bex, F., & Walton, D. (2016). Combining explanation and argumentation in dialogue. *Argument & Computation, 7*(1), 55–68.

Bibal, A., et al. (2021). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law, 29*, 149–169.

Biedermann, A., et al. (2020). Normative decision analysis in forensic science. *Artificial Intelligence and Law, 28*, 7–25.

Bourcier, D. (2003). *Legal knowledge and information systems: JURIX 2003: The sixteenth annual conference*. IOS Press.

Branting, L. K. (2003). A reduction-graph model of precedent in legal analysis. *Artificial Intelligence, 150*(1–2), 59–95.

Branting, L. K. (2017). Data-centric and logic-based models for automated legal problem solving. *Artificial Intelligence and Law, 25*, 5–27.

Branting, L. K., & Branting, L. K. (2000). GREBE: integrating rules and precedents for legal analysis. *Reasoning with Rules and Precedents: A Computational Model of Legal Analysis,* 63–109.

Branting, L. K., et al. (2021). Scalable and explainable legal prediction. *Artificial Intelligence and Law, 29*, 213–238.

Brożek, B., Furman, M., Jakubiec, M., & Kucharzyk, B. (2023). The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artificial Intelligence and Law,* 1–14.

Buchanan, B. G., & Shortliffe, E. H. (1984). Rule based expert systems: The MYCIN experiments of the Stanford heuristic programming project (the Addison-Wesley series in artificial intelligence), Addison-Wesley Longman Publishing Co., Inc.

Burgemeestre, B., et al. (2011). Value-based argumentation for justifying compliance. *Artificial Intelligence and Law, 19*, 149–186.

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

Caroprese, L., et al. (2022). Argumentation approaches for explanaible AI in medical informatics. *Intelligent Systems with Applications, 16*, 200109.

Chalkidis, I., et al. (2021). LexGLUE: A benchmark dataset for legal language understanding in English. arXiv preprint arXiv:2110.00976

Chalkidis, I., & Kampas, D. (2019). Deep learning in law: Early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law, 27*(2), 171–198.

Charwat, G., et al. (2015). Methods for solving reasoning problems in abstract argumentation—A survey. *Artificial Intelligence, 220*, 28–63.

Chen, D. L., & Eagel, J. (2017). Can machine learning help predict the outcome of asylum adjudications? *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law* (pp. 237–240).

Collenette, J., et al. (2020). An explainable approach to deducing outcomes in European Court of Human Rights cases using ADFs. *COMMA*.

Council of the European Union. (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts - presidency compromise text*. Accessed August 21, 2023, from https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf

Dadgostari, F., et al. (2021). Modeling law search as prediction. *Artificial Intelligence and Law, 29*, 3–34.

Dattachaudhuri, A., et al. (2021). A transparent rule-based expert system using neural network. *Soft Computing, 25*, 7731–7744.

Deeks, A. (2019). The judicial demand for explainable artificial intelligence. *Columbia Law Review, 119*(7), 1829–1850.

Devins, C., et al. (2017). The law and big data. *Cornell JL & Public Policy, 27*, 357.

Di Porto, F., & Zuppetta, M. (2021). Co-regulating algorithmic disclosure for digital platforms. *Policy and Society, 40*(2), 272–293.

Dung, P. M. (1995). An argumentation-theoretic foundation for logic programming. *The Journal of Logic Programming, 22*(2), 151–177.

Dunne, P. E., et al. (2011). Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence, 175*(2), 457–486.

El Ghosh, M., et al. (2017). Towards a legal rule-based system grounded on the integration of criminal domain ontology and rules. *Procedia Computer Science, 112*, 632–642.

Elston, D. M. (2019). Mendeley. *Journal of the American Academy of Dermatology, 81*(5), 1071.

Evans, T., et al. (2022). The explainability paradox: Challenges for xAI in digital pathology. *Future Generation Computer Systems, 133*, 281–296.

Górski, L., et al. (2021). Towards grad-cam based explainability in a legal text processing pipeline. arXiv preprint arXiv:2012.09603

Greenstein, S. (2022). Preserving the rule of law in the era of artificial intelligence (AI). *Artificial Intelligence and Law, 30*(3), 291–323.

Guidotti, R., et al. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR), 51*(5), 1–42.

Gutierrez, et al. (2023). A proposal for a definition of general purpose artificial intelligence systems. *DISO, 2*, 36.

Hafner, C. D., & Berman, D. H. (2002). The role of context in case-based legal reasoning: Teleological, temporal, and procedural. *Artificial Intelligence and Law, 10*(1–3), 19–64.

Hamdani, R. E., et al. (2021). A combined rule-based and machine learning approach for automated GDPR compliance checking. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*.

Heras, S., et al. (2009). Challenges for a CBR framework for argumentation in open MAS. *The Knowledge Engineering Review, 24*(4), 327–352.

Horty, J. F., & Bench-Capon, T. J. (2012). A factor-based definition of precedential constraint. *Artificial Intelligence and Law, 20*, 181–214.

Islam, M. B., & Governatori, G. (2018). RuleRS: A rule-based architecture for decision support systems. *Artificial Intelligence and Law, 26*(4), 315–344.

Islam, M. R., et al. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences, 12*(3), 1353.

Keane, M. T., & Kenny, E. M. (2019). *How case-based reasoning explains neural networks: A theoretical analysis of XAI using post-hoc explanation-by-example from a survey of ANN-CBR twin-systems*. Springer.

Kliegr, T., et al. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence, 295*, 103458.

Kuiper, O., van den Berg, M., van der Burgt, J., & Leijnen, S. (2022). Exploring explainable AI in the financial sector: Perspectives of banks and supervisory authorities. *Artificial Intelligence and Machine Learning: 33rd Benelux Conference on Artificial Intelligence, BNAIC/Benelearn 2021, Esch-sur-Alzette, Luxembourg, November 10–12, 2021, Revised Selected Papers 33* (pp. 105–119). Springer International Publishing.

Kökciyan, N., et al. (2017). An argumentation approach for resolving privacy disputes in online social networks. *ACM Transactions on Internet Technology (TOIT), 17*(3), 1–22.

Labrie, N., & Schulz, P. J. (2014). Does argumentation matter? A systematic literature review on the role of argumentation in doctor– patient communication. *Health Communication, 29*(10), 996–1008.

Lamy, J.-B., et al. (2019). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine, 94*, 42–53.

Langer, M., et al. (2021). What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*, 103473.

Liepiņa, R., et al. (2020). Arguing about causes in law: A semi-formal framework for causal arguments. *Artificial Intelligence and Law, 28*(1), 69–89.

Lippi, M., et al. (2019). CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law, 27*, 117–139.

Liu, Q., et al. (2021). Towards an efficient rule-based framework for legal reasoning. *Knowledge-Based Systems, 224*, 107082.

Longo, L., et al. (2020). *Explainable artificial intelligence: Concepts, applications, research challenges and visions*. Springer.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Marques Martins, J. (2020). A system of communication rules for justifying and explaining beliefs about facts in civil trials. *Artificial Intelligence and Law, 28*, 135–150.

Matulionyte, R., & Hanif, A. (2021). A call for more explainable AI in law enforcement. *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*. IEEE.

McCarty, L. T. (1976). Reflections on TAXMAN: An experiment in artificial intelligence and legal reasoning. *Harv. L. Rev., 90*, 837.

Mehrabi, N., et al. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), 1–35.

Modgil, S., & Prakken, H. (2014). The ASPIC+ framework for structured argumentation: A tutorial. *Argument & Computation, 5*(1), 31–62.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264–269.

Mowbray, A., et al. (2023). Explainable AI (XAI) in Rules as Code (RaC): The DataLex approach. *Computer Law & Security Review, 48*, 105771.

Možina, M., et al. (2007). Argument based machine learning. *Artificial Intelligence, 171*(10–15), 922–937.

Muddamsetty, S. M., et al. (2022). Visual explanation of black-box model: Similarity Difference and Uniqueness (SIDU) method. *Pattern Recognition, 127*, 108604.

Neil, M., et al. (2019). Modelling competing legal arguments using Bayesian model comparison and averaging. *Artificial Intelligence and Law, 27*, 403–430.

Palmer, H., & Cohen, K. (2022). Genetic fuzzy hand gesture classifier. *Explainable AI and other applications of fuzzy techniques: Proceedings of the 2021 Annual Conference of the North American Fuzzy Information Processing Society, NAFIPS 2021* (pp. 332–342). Springer International Publishing.

Palmirani, M. et al. (2012). AI Approaches to the Complexity of Legal Systems—Models and Ethical Challenges for Legal Systems, Legal Language and Legal Ontologies, Argumentation and Software Agents: International Workshop AICOL-III, Held as Part of the 25th IVR Congress, Frankfurt am Main, Germany, August 15-16, 2011. Revised Selected Papers, Springer.

Prakken, H. (2010). An abstract framework for argumentation with structured arguments. *Argument & Computation, 1*(2), 93–124.

Prakken, H., et al. (2015). A formalization of argumentation schemes for legal case-based reasoning in ASPIC+. *Journal of Logic and Computation, 25*(5), 1141–1166.

Ribeiro, M. T., et al. (2016). Why should i trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Richmond, K. M. (2021). AI, machine learning, and international criminal investigations: The lessons from forensic science. *Retskraft: Copenhagen Journal of Legal Studies, 5*(1), 31–58.

Rissland, E. L., & Skalak, D. B. (1989). Combining case-based and rule-based reasoning: A heuristic approach. *IJCAI*.

Rissland, E. L., & Skalak, D. B. (1991). CABARET: Rule interpretation in a hybrid architecture. *International Journal of Man-Machine Studies, 34*(6), 839–887.

Rissland, E. L., Skalak, D. B., & Friedman, M. T. (1996). BankXX: Supporting legal arguments through heuristic retrieval. *Artificial Intelligence and Law, 4*(1), 1–71.

Rissland, E. L., et al. (2003). AI and Law: A fruitful synergy. *Artificial Intelligence, 150*(1–2), 1–15.

Rissland, E. L., et al. (2005). Case-based reasoning and law. *The Knowledge Engineering Review, 20*(3), 293–298.

Rosengrün, S. (2022). Why AI is a threat to the rule of law. *DISO, 1*, 10.

Rotolo, A., & Sartor, G. (2023). AI & Law: Case-based reasoning and machine learning. In *Encyclopedia of the Philosophy of Law and Social Philosophy*, 1–7

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215.

Schulz, C., & Toni, F. (2016). Justifying answer sets using argumentation. *Theory and Practice of Logic Programming, 16*(1), 59–110.

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 1–59.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Šešelja, D., & Straßer, C. (2013). Abstract argumentation and explanation applied to scientific debates. *Synthese, 190*, 2195–2217.

Steging, C., Renooij, S., & Verheij, B. (2021). Discovering the rationale of decisions: towards a method for aligning learning and reasoning. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (pp. 235–239).

Sørmo, F., et al. (2005). Explanation in case-based reasoning—Perspectives and goals. *Artificial Intelligence Review, 24*, 109–143.

Surden, H. (2019). Artificial intelligence and law: An overview. *Georgia State University Law Review, 35*, 19–22.

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813.

Toulmin, S. (1958). *The uses of argument*. Cambridge: University Press.

Vassiliades, A., et al. (2021). Argumentation and explainable artificial intelligence: A survey. *The Knowledge Engineering Review, 36*, e5.

Verheij, B. (2003). Artificial argument assistants for defeasible argumentation. *Artificial Intelligence, 150*(1–2), 291–324.

Verheij, B. (2016). Formalizing value-guided argumentation for ethical systems design. *Artificial Intelligence and Law, 24*, 387–407.

Verheij, B. (2017). Proof with and without probabilities: Correct evidential reasoning with presumptive arguments, coherent hypotheses and degrees of uncertainty. *Artificial Intelligence and Law, 25*, 127–154.

Vlek, C. S., et al. (2014). Building Bayesian networks for legal evidence with narratives: A case study evaluation. *Artificial Intelligence and Law, 22*, 375–421.

Vlek, C. S., et al. (2016). A method for explaining Bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law, 24*, 285–324.

Wachter, S., et al. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review, 41*, 105567.

Walton, D. (2019). When expert opinion evidence goes wrong. *Artificial Intelligence and Law, 27*(4), 369–401.

Wyner, A. (2008). An ontology in OWL for legal case-based reasoning. *Artificial Intelligence and Law, 16*, 361–387.

Wyner, A., & Bench-Capon, T. (2007). Argument schemes for legal case-based reasoning. *JURIX*.
Wyner, A. Z., et al. (2011). Towards formalising argumentation about legal cases. *Proceedings of the 13th International Conference On Artificial Intelligence and Law*.
Yu, S., & Chen, X. (2023). How to justify a backing's eligibility for a warrant: The justification of a legal interpretation in a hard case. *Artificial Intelligence and Law, 31*(2), 239–268.
Zhang, Y., et al. (2022). Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics, 12*(2), 237.
Zheng, H., et al. (2021). Logical comparison of cases. AI Approaches to the Complexity of Legal Systems XI-XII: AICOL International Workshops 2018 and 2020: AICOL-XI@ JURIX 2018, AICOL-XII@ JURIX 2020, XAILA@ JURIX 2020, Revised Selected Papers XII, Springer.
Zhong, Q., et al. (2019). An explainable multi-attribute decision model based on argumentation. *Expert Systems with Applications, 117*, 42–61.
Zhou, B., et al. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Authors and Affiliations

Karen McGregor Richmond[1] · Satya M. Muddamsetty[2] · Thomas Gammeltoft-Hansen[1] · Henrik Palmer Olsen[1] · Thomas B. Moeslund[2]

✉ Karen McGregor Richmond
karen.richmond@jur.ku.dk

Satya M. Muddamsetty
smmu@create.aau.dk

[1] Faculty of Law, Copenhagen University, Karen Blixen's Plads 16, Copenhagen 2400, Denmark

[2] Department of Architecture, Design, and Media Technology (CREATE), Visual Analysis and Perception Lab, Aalborg University, Rendsburggade 14, Aalborg 9000, Denmark