Artificial Intelligence 289 (2020) 103387

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Explanation in AI and law: Past, present and future ☆

Katie Atkinson *, Trevor Bench-Capon, Danushka Bollegala

Department of Computer Science, University of Liverpool, Liverpool, UK

ARTICLE INFO

Article history:
Received 28 February 2020
Received in revised form 3 September 2020
Accepted 12 September 2020
Available online 16 September 2020

Keywords:
Explainable AI
AI and law
Computational models of argument
Case-based reasoning

ABSTRACT

Explanation has been a central feature of AI systems for legal reasoning since their inception. Recently, the topic of explanation of decisions has taken on a new urgency, throughout AI in general, with the increasing deployment of AI tools and the need for lay users to be able to place trust in the decisions that the support tools are recommending. This paper provides a comprehensive review of the variety of techniques for explanation that have been developed in AI and Law. We summarise the early contributions and how these have since developed. We describe a number of notable current methods for automated explanation of legal reasoning and we also highlight gaps that must be addressed by future systems to ensure that accurate, trustworthy, unbiased decision support can be provided to legal professionals. We believe that insights from AI and Law, where explanation has long been a concern, may provide useful pointers for future development of explainable AI.

1. Introduction

The English essayist Charles Lamb famously wrote "*He is no lawyer who cannot take two sides*". For many, the same is true of AI and Law programs. Arguing for just one side, or worse simply pronouncing for one side, is not enough. To make a convincing case in court, one must be able to offer reasons for one's own side, and to anticipate and rebut arguments for the other side. The first important AI and Law program, TAXMAN [86], set out to reconstruct the arguments of the majority and minority opinions in the famous tax case of *Eisner v Macomber*[1] In TAXMAN, there was no interest in assessing and deciding between the two opinions: the purpose was simply to be able to argue for both sides. The point is that the outcome of a case is often not clear: in any serious legal dispute there are opposing arguments, and very often opinions differ as to who has the better of it. Decisions are reversed on appeal, and may be reversed again at the highest level of appeal. Even at the highest level, where the most gifted lawyers are the judges, consensus is, as in *Eisner*, far from invariable: an article in the *Washington Post*[2] stated

> "According to the Supreme Court Database, since 2000 a unanimous decision has been more likely than any other result — averaging 36 percent of all decisions. Even when the court did not reach a unanimous judgment, the justices often

☆ This paper is part of the Special Issue on Explainable AI.

* Corresponding author.
E-mail address: K.M.Atkinson@liverpool.ac.uk (K. Atkinson).

[1] Eisner v Macomber, 252 U.S. 189 (1920) was a case concerning the tax liability of a stock dividend paid in the form of additional shares. The majority (5-4) found in favour of Mrs Macomber.

[2] https://www.washingtonpost.com/news/posteverything/wp/2018/06/28/those-5-4-decisions-on-the-supreme-court-9-0-is-far-more-common/ Accessed 13th November 2019.

https://doi.org/10.1016/j.artint.2020.103387
0004-3702/

secured overwhelming majorities, with 7-to-2 or 8-to-1 judgments making up about 15 percent of decisions. The 5-to-4 decisions, by comparison, occurred in 19 percent of cases".

Although this article was in fact arguing that consensus was the norm, the results still indicate disagreement in the significant majority of cases, and the narrowest of majorities in nearly a fifth of cases. Consensus may be the most likely result of the ten possible (the quorum is 8), but disagreement remains far more likely, and 5-4 the second most likely result. Given that even the most expert people can disagree, it would not be reasonable to accept a judgement from a machine unless backed up with convincing reasons.

In recent years there has been some research directed towards the prediction of case outcomes using algorithms applied to large data sets (e.g. [5] and [88]), but for most of its history AI and Law has been far more interested in modelling the reasoning to explain the outcome (and to offer reasons for alternative possible outcomes) than in predicting the outcome itself. AI and Law therefore offers an interesting area in which to explore methods for the explanation of AI programs,[3] as advocated in the most recent Presidential Address to the International Association for AI and Law [129]. In this paper we will review a number of approaches. Before we do so, however, we will consider some general points about explanation, especially in law.

1.1. Right to explanation

Apart from the centrality of argumentation to legal reasoning, the intellectual challenge of modelling legal reasoning and the availability of, in the form of opinions on cases, a large volume of examples, there is another important reason why explanation is vital for artificial intelligence applied to law. This is the *right to explanation* [54]. In a legal dispute there will be two parties and one will win and one will lose. If justice is to be served, the losers have a right to an explanation of why their case was unsuccessful. Given such an explanation, the losers may be satisfied and accept the decision, or may consider if there are grounds to appeal. Justice must not only be done, but must be seen to be done, and, without an explanation, the required transparency is missing. Therefore explanation is essential for any legal application that is to be used in a practical setting.

1.2. Nature of explanation

In his recent illuminating survey [90], Miller gives four main findings of features of explanation. These are:

- Explanations are *contrastive*. As well as explaining why a particular classification is appropriate, a good explanation will also say why other classifications are not. This is often done using counterfactuals and hypotheticals: "*if x had been true, then the classification would have been A, not B*".
- Explanations are *selective*. Rarely is a logically complete explanation provided, but rather only the most salient points are presented unless more detail is required by the recipient of the explanation. The assumption is that there will be a considerable degree of shared background knowledge, and so the explanation need only point to some fact or rule as yet unknown to the recipient.
- Explanations are rarely in terms of *probabilities*. Using statistical generalisations to explain why events occur is unsatisfying since they do not explain the generalisation itself. Moreover, the explanation typically applies to a single case, and so would require some explanation of why that particular case is typical.
- Explanations are *social*. Explanations involve a transfer of knowledge, between particular people in a particular situation and so are relative to the explainer's beliefs about the explainee's beliefs.

Miller says that he believes "most research and practitioners in artificial intelligence are currently unaware" of these features. AI and Law, however, has long recognised these features, and made them an important part of its approach to explanation.

- Contrastive explanations can be found in legal case based systems such as HYPO ([108] and [7]), quite possibly the most influential of all AI and Law programs [21]. Indeed the name HYPO is itself short for *hypothetical*: one of the main motivations of the system was to explore how the hypothetical variations on cases would change their outcome. Also there are explanations based on the weighing of pro and con reasons such as the Reason Based Logic of Hage [69] or the tool developed by Lauritsen [77].
- Selective explanations were pursued by several AI and Law researchers. Often this was done through the use of argumentation schemes such as that of Toulmin [123], used in, for example, [83] and [26]. The idea was to present the key data items which gave rise to the inference and to suppress things that should be expected to be already known, such as "John is a man" or "67 > 65", unless explicitly requested by the user.

[3] The lack of effective explanations is currently a matter of concern for potential end users of AI: in a 2019 global survey *From Roadblock to Scale: The Global Sprint Towards AI*, commissioned by IBM, 83% of global respondents felt that "being able to explain how AI arrived at a decision is universally important". See: http://filecache.mediaroom.com/mr5mr_ibmnews/183710/Roadblock-to-Scale-exec-summary.pdf (last accessed 20th July 2020).

- Probabilities are rarely used in legal decisions. Even where Bayesian reasoning is used in AI and Law, the explanation is presented not in probabilistic terms but as scenarios [38], [130] or arguments [121]. A legal decision is supposed to determine what is true on the facts of the particular case. An 80% probability would mean that one in five cases would be decided wrongly, which would not be justice.
- Legal explanations are inherently social, occurring in the context of courtroom procedure, and involving an interaction between plaintiff, defendant, judge and, possibly, jury. This is reflected in the popularity of dialogues as the vehicle of explanation in AI and Law, such as [71], [61], [16] and [128].

Thus AI and Law provides an excellent domain in which to study explanation of AI systems. In AI and Law explanation has a long history, is a mandatory feature of fielded applications in the legal domain, and AI and Law has long recognised the important facets of explanation identified in [90]. The rest of this paper will be structured as follows. Section 2 will give an overview of the main types of explanation used in AI and Law. The various types of explanation will then be described in more detail in sections 3, 4 and 5. Section 6 will look at interactive explanations, section 7 will consider efforts to explain machine learning in AI and Law and section 8 will look at some current research directions that may become influential in the future. Concluding remarks are provided in section 9.

2. Explanation in AI and law

In its early days, AI and Law followed two main approaches: cased based approaches such as TAXMAN [86], HYPO [108] and CATO [6], and rule based approaches including Gardner's account [57], approaches using production rules [114], and approaches using logic programming [115]. For some time case and rule based reasoning were seen as alternatives [35]. Each gave rise to distinctive styles of explanation: case based systems tended to explain by offering precedent cases as examples, while rule based approaches could offer a trace of the inference process in the manner of classic expert systems such as MYCIN [46].

2.1. Explanation by example

The idea of using examples for explanation was pioneered by Rissland in [106] and [107]. Rissland was originally inspired by the work in mathematics of Lakatos [76], but soon realised that this technique was also applicable to law [106]. In the Common Law tradition of the United States, lawyers typically argue by citing precedent cases which favour their side and distinguishing precedent cases which favour the other side. Explanation therefore tends to take the form: *the case should be decided in this way because it is like these cases, and unlike these other cases*, a form of contrastive explanation, making use of both positive and negative examples. Another idea motivating [108] was that the explanation would be enhanced by citing hypothetical features of the case which, had they been different, would have changed the outcome, by making it sufficiently like an adverse precedent. Example based approaches will be described in detail in section 3.

2.2. Explanation using rules

Although cases are a strong feature of Common Law traditions, laws are paradigmatically found in statutes. This is especially true of the European tradition of Civil Law. In many areas, these statutes can be seen as offering definitions of particular concepts such as murder, benefit entitlement and citizenship. This was exploited in [115] which provided a logical formalisation of a piece of legislation which could then be executed as a logic program. This approach proved highly influential and inspired several other researchers, including [117] and [73].

As well as logic programs, typically based on formalisations of legislation, there were some more traditional expert systems, using production rules and based on knowledge elicited from a domain expert such as [114] and [119]. In these systems the rules often represent sufficient conditions taken from cases and commentaries rather than definitions taken from statute. All types of rule based system, however, offered their explanations in the standard expert systems form of the *how*, *why* and *what-if* explanations pioneered by MYCIN [46].

2.3. Hybrid systems

Some approaches attempted to combine rule and case based reasoning, although assigning them different roles. The idea was that the law could be described at a high level in terms of rules, but that determining whether these rules were satisfied by the facts of a particular case required case based reasoning. This idea was pioneered by Skalak and Rissland [118] and further developed by Brüninghaus and Ashley [45]. Rule based and hybrid systems will be discussed in section 4.

2.4. Explanation with reasons: argumentation

Although all the above AI and Law systems modelled arguments, the explicit use of computational argumentation has become increasingly popular in AI and Law. Two developments are particularly significant here: the development of abstract argumentation [55] and the use of argumentation schemes. At first only the scheme of Toulmin [123] was used, but later

a variety of schemes as advocated by Walton [131] were explored, e.g. [127] and [100]. The use of argumentation for explanation in AI and Law will be discussed in section 5.

2.5. Example cases used in this paper

In the following sections we will consider representative examples of the aforementioned approaches in detail. We will use a running example based on cases involving the ownership of wild animals (eventually extended to include *Popov v Hayashi*, which concerned a disputed baseball [138]). These cases often form part of the introduction to property law in US Law Schools, and were introduced into AI and Law in [36], since when they have been widely discussed by a variety of different researchers. A special issue of AI and Law journal considered different approaches to modelling these cases [9]. Although a number of cases have been discussed in the literature, we will focus on the three cases used in [36], which feature in all such discussions. We summarise the cases to enable appreciation of the differences in the various explanation methods provided later in the paper.

- *Keeble v Hickergill (1707)*. This was an English case in which Keeble rented a duck pond, to which he lured ducks, which he shot and sold for consumption. Hickergill, out of malice, scared the ducks away by firing guns. The court found for Keeble. Two arguments for Keeble are possible: that he was engaged in an economically valuable activity, and that he was operating on his own land. The former reading is adopted in [36], but others, e.g. [31], prefer the latter.
- *Pierson v Post (1805)*. In this New York case, Post (the plaintiff at first instance) was hunting a fox with hounds. Pierson intercepted the fox, killed it with a handy fence rail, and carried it off. The court found for Pierson. The argument was that Post had never had possession of the fox. The argument that hunting vermin is a useful activity which needs protection and encouragement formed the basis of the minority opinion.
- *Young v Hitchens (1844)*. In this English case, Young was a commercial fisherman who spread a net of 140 fathoms in open water. When the net was almost closed, Hitchens went through the gap, spread his net and caught the trapped fish. The case was decided for Hitchens. The basis for this was that Young had never had possession of the fish, and that it was not part of the court's remit to rule as to what constituted unfair competition.

In our examples we will take *Young* as the case under consideration and *Pierson* and *Keeble* as the precedents.

3. Explanation through examples: case based reasoning

Although there have been several approaches to reasoning with legal cases, including the use of prototypes and deformations [87] and semantic networks [40], by far the dominant approach has been the use of dimensions and factors [21]. This approach will therefore be the one considered in detail in this section.

3.1. Dimensions and factors

A basic principle of common law is that like cases should be treated in a like manner, embodied in the notion of *stare decisis* ("let the decision stand"); this says that like cases should be decided in the same way, that previous cases provide precedents to be followed unless there is a good reason not to, i.e. the current case can be distinguished from the precedent in some significant way. This raises the question of how it can be determined whether two cases are sufficiently similar. The facts of the cases are always rather particular, and may look, at first sight, rather disparate, as we now illustrate.

The famous series of negligence cases discussed by Levi in [79] has cases involving a loaded gun, a possibly defective gun, mislabelled poison, defective hair wash, scaffolds, a defective coffee urn, a defective aerated bottle, a defective carriage, a bursting lamp, a defective balance wheel for a circular saw, and a defective boiler. In the decisions the items up to the aerated bottle are considered *like*, and the remainder *unlike*. This is because the various objects were not considered as objects in their usual sense, but according to an attribute they possessed which had legal significance in the particular situation, namely whether they were *imminently dangerous*, enabling the list of items to be split into categories of 'imminently dangerous' and 'latently dangerous'. Here *imminently* and *latently* dangerous are essentially legal notions, established through the series of cases. The argument is settled by examining the reasons for attribution in previous cases, not by asking the opinion of native English speakers. This explains why negligence was found for the first set of items but not the second: these items were imminently dangerous. The explanation does not turn on the particular facts but rather on the legal concept that those facts support.

To capture the above, HYPO [108] and [7], developed the notion of a *dimension*, a legally significant aspect of a case representing a range of values starting from the point most favourable to the plaintiff at one end, and then increasingly favouring the defendant until the point most favourable to the defendant is reached at the other end. The facts of the case determine where it should be positioned on the dimension. In the cases above, the imminence of the danger would be a dimension, and the various items arranged along it with, perhaps the defective coffee urn the most imminently dangerous and the defective but unloaded gun the least imminently dangerous of those favourable to the plaintiff. The pro-defendant items could also be arranged similarly. Now for a case with a non-defective unloaded gun, one could find for the defendant

by explaining that the gun was not imminently dangerous, although if it had been loaded or defective, the case would have been found for the plaintiff, a sort of contrastive explanation.

In [6] dimensions were replaced by *factors*. Factors eliminate the notion of degree inherent to dimensions so that factors are always simply present or absent in a case, and always favour one of the two sides. One way of viewing them is a point or ranges on a dimension, and with a fixed crossover point from plaintiff to defendant imposed [103]. In [6], cases are represented directly as sets of factors, rather than as sets of facts. The focus in [6] is on *distinguishing*: what makes one case significantly different from another. Factors proved popular and formed the basis of most approaches to reasoning with legal cases until a recent revival of interest in dimensions [23], [72] and [105]. For discussions of the relationship between dimensions and factors see [31] and [109].

The explanation in these systems takes the form of a three-ply dialogue. First the plaintiff cites a case which has the most similarity to the current case of the precedents favouring the plaintiff. Then the defendant replies by offering counter examples, precedents found for the defendant at least as similar as the case cited for the plaintiff, and by distinguishing the cited case. Finally the plaintiff attempts a rebuttal, distinguishing the counter examples, and offering reasons why the distinctions are not significant. The case for the defendant can be made by reversing the roles in the three-ply structure. Explanations are thus essentially through the presentation of examples, and can be seen as contrastive, selective and social.

3.2. Application to the Wild Animals example

Let us now apply this approach to the Wild Animals example introduced above. We will suppose *Keeble* and *Pierson* to be precedents and *Young* the case under consideration. There are many representations of these cases in terms of dimensions and factors to choose from, but we will largely follow [109]. They began by listing four dimensions:

D1 (Control/Possession) concerns control and possession of the game by the hunter.
D2 (Site) concerns whether the site where the game was taken or pursued is characterized as public land or private land of the hunter.
D3 (Livelihood) concerns whether the hunter was pursuing the game in order to make his livelihood or for sport.
D4 (Competition) concerns the possibility of there being economic competition between the plaintiff and the defendant.

These could form the basis of the factors in [36]. We can simplify D1 into *F1 NotCaught*, covering all the values of D1 favouring the defendant; Next we split D2 into *F2a Private* and *F2b Public*. This is, of course, a simplification of several kinds of tenure, but the suggestion is that it is whether the land can be seen as the plaintiff's own that matters. Similarly D3 can be treated as Boolean, *F3 EarningLivelihood*, as can D4, *F4 competition*, simply favouring the defendant if he is in competition with the plaintiff.

The cases can now be represented as sets of factors:

Keeble : F1, F2a, F3
Pierson : F1, F2b
Young : F1, F2b, F3, F4

Suppose now we wish to offer *Keeble* as an example favouring the plaintiff in *Young*. We can say:

Where: Plaintiff had not caught the game (F1), and Plaintiff makes his livelihood from taking game (F3), Plaintiff should win claim. Cite: Keeble.

But this can be rebutted because there are two distinctions so that:

Keeble is distinguishable because: In Keeble, the game area is plaintiff's property (F2a). This is not so in Young. In Young, plaintiff and defendant compete (F4). This was not so in Keeble.

The defendant can now offer Pierson as his own precedent to provide a counterexample:

Where: the game is not under plaintiff's control (F1) and the game area is open (F2), defendant should win claim. Cite: Pierson.

The plaintiff may now attempt to distinguish Pierson, by saying:

Pierson is distinguishable because: In Pierson, the plaintiff was not earning his living (F3 absent). This is not so in Young.

This, however, can be rebutted, by claiming that F3 is not significant in this case as its effect is cancelled by the presence of F4.

On this basis, a finding for the defendant seems plausible. Note, however, these systems do not come to a conclusion as to what the decision should be. They explain why one might find for the plaintiff and why one might find for the defendant, and allow the user to choose. Thus in *Young*, the user decides whether F4 is enough to render F3 insignificant when distinguishing *Young* from *Pierson*, and, if so, whether the presence of F3 is alone sufficient to distinguish *Pierson* or whether F2a is also needed.

3.3. Explanation with dimensions and magnitudes

Recently there has been a good deal of interest in returning to dimensions or factors with magnitudes, e.g. [72] and [105]. This enables explanation to be given in terms of the weighing of pro and con reasons, as found in the Reason Based Logic of Hage [69]. A tool for visual exploration of using different weights for different factors is described in [77]. Also, a threshold can be set, so that a factor must be present to a sufficient extent to be deemed worthy of consideration [24]. As well as balancing sets of reasons for and against a decision, the explanation can be given in terms of a trade-off between factors. For example privacy and the urgency required for law enforcement in cases relating to the automobile exception of the US Fourth Amendment [30]. Precedents can set limits of the degree of trade off permitted, as shown graphically in [23].

4. Step by step explanation: rule based reasoning

For our example of this style of system we will consider logic programming in the style of [115], or, as applied to case law, [33]. For case law, this approach requires that a set of rules be derived from the precedents, encapsulating the knowledge that they represent. This does, however, require some degree of interpretation on the part of a knowledge engineer or domain expert. Moreover, the interpretation is subject to change, and the rules may require reconsideration in the light of new cases (see [79], [51] and [27]).

We begin by laying down some background knowledge: that ownership of a wild animal may be established either by owning the land on which it is to be found, or by taking possession of it through capture.

This gives:

R1 findFor(plaintiff) if capture.
R2 findFor(plaintiff) if ownLand.

We must now define these two concepts. From *Pierson* we learn from the majority opinion that bodily possession is certainly sufficient for capture, and having the animal within one's control may be sufficient. Since in *Pierson* the plaintiff had neither captured the animal nor gained control of it, the issue requires that we make an interpretation. Reading the decision suggests that the stricter position was advocated by the majority opinion, as so we adopt this.

R3 capture if bodilyPossession.

Turning now to *Keeble*, we can see that renting the land is sufficient to establish ownership. *A fortiori*, actual ownership is also sufficient.

R4a ownLand if owned.
R4b ownLand if rented.

In [36], however, the authors argue that Keeble could also win on capture because he was in control of the ducks – if not scared away he could shoot them when he pleased – and was earning his livelihood.

R3a capture if control and livelihood.

This is consistent with our interpretation of *Pierson* in R3, since we have the extra condition. Testing these rules with the query findFor(plaintiff), we get *no* for *Pierson* and *true* for *Keeble*. We get no explanation for *Pierson*. One weakness of this approach was that the explanation of negatives was not straightforward [28]: the standard *how* explanation will state how something has been proved, not how things failed to be proved. For *Keeble*, however, we will get an explanation:

I can show *findFor(plaintiff)* because I can show *ownLand*.
I can show *ownLand* because I can show *rented*.

We also have a second explanation, based on capture:

I can show *findFor(plaintiff)* because I can show *capture*.
I can show *capture* because I can show *control* and *livelihood*.

If we now use these rules to determine the outcome in *Young*, we will get the answer to the query findFor(plaintiff) as *true*. If we know the actual outcome was for the defendant, we will want that explained. This will be the second explanation of *Keeble*. Confronted with this, however, an astute defence counsel will note that the fact of the defendant being in competition with the plaintiff has not been used, and so argue that R3a should not be followed in this case. This argument was successful, and so R3a should be modified to:

R3b capture if control and livelihood and not(competition).

This reinterpretation of an existing rule in the light of a new case, fits the mechanism for dynamic case law described in [79] and [27]. With this modification the rules now find correctly for *Young*, while still finding correctly for *Keeble*, where there was no competition (and also rule 4b applies). However, there is no explanation for the outcome for the defendant in *Young*: this outcome is the default which holds when the plaintiff is unable to satisfy either of the conditions.

Compared with the case based approach, there is more effort required to build the rule based system. Although both approaches require the identification of the relevant factors/predicates, the rule based approach has the additional burden of interpreting the cases to provide a set of rules. Note also, that when the case contains a new factor, as with *competition* in *Young*, which can be used to distinguish the case in the case based approach, the rules will need to be questioned. Every case has the opportunity of modifying or extending the rule base [27].

4.1. Conditional answers and multiple solutions

A major difficulty of this style of explanation is that only positive answers can be explained. In a classical expert system such as MYCIN [46] this was less of a problem. Firstly such expert systems were often termed *consultative*: the system was supposed to have the requisite information and the user was supposed to accept the answers, whereas in law we have the *right to explanation*. Secondly, in MYCIN, the knowledge itself is more stable: the human body is not subject to radical change, whereas a legal case always has the power to change the existing wisdom. Thirdly, MYCIN had a large number of options, so justifying the chosen answer was more sensible, since there was not a single alternative to explain away. Fourthly, the right to explanation in law [54] means that the losing party is even more interested in the explanation than the winner.

Some of what was required could be achieved by the use of the *what-if* query. For example, one could ask of *Young what if competition had been false?*. This, however, required the user to form a hypothesis about why the plaintiff had failed. One approach proposed to meet these difficulties was the conditional answer approach of [136]. Here the system suggested ways in which the desired outcome could be made true. So if the plaintiff in *Young* were to use the system, he would be told that he would win provided he had captured the animals (R1). Stepping down he would find that this could be shown provided that he had been in bodilyPossession (R3). Since this was not so, an alternative would be sought, and he would be told that, since he was in control of the animals and pursuing his livelihood, he would win provided the defendant was not in competition with him (R5a). Again seeking an alternative, he would be told that he could win if it was his own land. Since the incident took place on the high seas, Young will give up here. He will, however, have a thorough understanding of why he did not win. Note that this explanation has elements of contrastive explanation and of selective explanation, since the users abandon a line of enquiry once they are satisfied that it is of no use to them, as when Young was well aware he was not on his own land.

An approach to problems with the need to make an interpretation was proposed in [33] and [113]. The idea here was that instead of deciding on a single set of rules, intended to give a single, putatively definitive, answer, all plausible rules, together with their source should be represented. For example in addition to the rules above we could include:

[R3c, [PiersonvPost, minority]] capture if hotPursuit and usefulActivity

to represent the minority opinion of Livinston in *Pierson* that the bodily possession requirement should be relaxed to encourage the socially useful activity of hunting vermin. Now this possibility is indicated when the query is run against the facts of *Pierson*. Although the rule was rejected when *Pierson* was heard, it might be that social values have since changed and the rule could therefore be acceptable to some future court.

This approach thus presents a variety of possibilities which the user must choose between. Note that here there is no single answer, and no single explanation. Rather a range of explanations for different outcomes are presented and the user invited to choose between them. The strategy, as expressed in [33] was:

> In applications where we require legal decision support we have proposed a system of conflicting rules. These rules are designed to present the relevant arguments for and against the conclusion as a basis on which the user can make his own decision. In the law, questions of open texture are resolved by the presentation of a case before a judge. The judgement will be a reasoned decision to accept an argument.

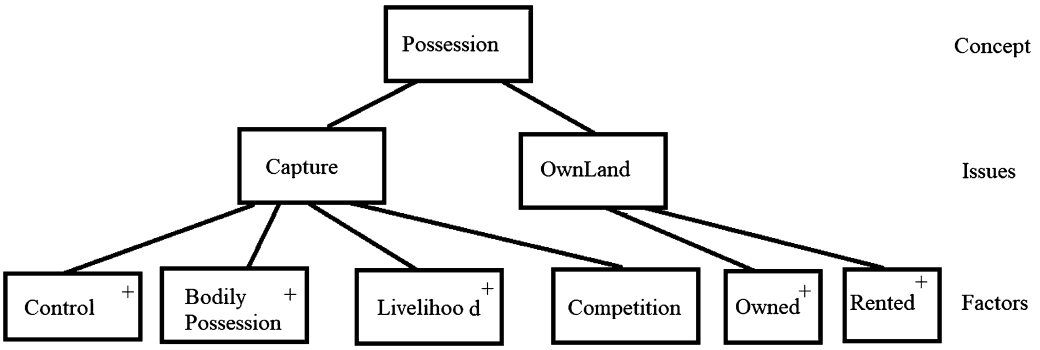

Fig. 1. Abstract factor hierarchy for wild animals.

4.2. Hybrid systems

The rule based approach had the advantage of structuring the explanation according to the underlying statute or legal doctrine, but tended to be rather prescriptive and required considerable knowledge engineering effort in constructing the rule base. In order to try to get the best of both worlds, some researchers developed hybrid systems. Examples of such approaches are CABARET [118] and IBP [45].

The idea here was the domain would be described at a high level as a set of general rules (termed a *logical model* in [45]). In our example this would comprise R1 and R2, representing the two routes to establish possession. Factors would now be grouped into a hierarchy (the *abstract factor hierarchy* of CATO [6]). Factors are marked "+" or "-" to show whether they support or oppose the presence of their parent. The hierarchy for our example is shown in Fig. 1. Now explanation of whether or not the leaf nodes of the logical model are satisfied can be given in terms of the cased based reasoning of CATO, although only the factors relevant to the particular issue would be included, thus focusing attention of the aspects relevant to the issue under consideration.

Now the explanation is presented on an issue by issue basis. First capture is considered:

> Where: Plaintiff had not caught the game (F1), and Plaintiff makes his livelihood from taking game (F3), Plaintiff should win claim on grounds of capture. Cite: Keeble.

But here the rebuttal, unlike the pure case based system, does not mention land ownership:

> Plaintiff should not win claim on grounds of capture. Keeble is distinguishable because: In Young, plaintiff and defendant compete (F4). This was not so in Keeble.

On the land ownership issue, the plaintiff has no argument, whereas the defendant does:

> Where: The game area was not private (F2a absent), Plaintiff should not win claim on grounds of land ownership. Cite: Pierson.

We will therefore find both issues for the defendant, as was the case in *Young*. Where some issues favour one side and others favour the other, as would happen if we had a case with the facts of *Keeble*, but with the defendant in competition with the plaintiff, then we need to rank the issues. What is needed there is to establish that land ownership has priority over capture. One solution to this was to use the underlying purposes or values of the law. We will discuss this in the next section.

4.3. Values

In the case based and hybrid approaches and some rule based approaches, such as the multiple solutions approach, the system presents options but does not offer reasons for choosing between them. Thus, for example, whether Young wins depends on whether it is accepted that being in competition with Young is sufficient to justify the defendant's interference in Young's pursuit. So the question arises: on what should these choices be based? An answer was offered in [36], which suggested that the answer could be determined by a consideration of the social purposes of the law, and which decision would serve these purposes better. In *Pierson* they argue that the choice is between the clarity (and consequently the reduced litigation) that will arise from requiring the very determinate criterion of bodily possession as against the purpose of encouraging the socially useful activity of hunting vermin that would result from a vaguer criterion. In *Keeble*, in contrast, the more relaxed criterion would encourage the *economically* useful activity of supplying ducks to the marketplace. Accord-

ing to [36], this would justify additional litigation where livelihoods were threatened. In *Young*, the competition means that economically it does not matter who lands the fish, and so the court followed *Pierson*.

This idea was developed in [32], renaming the purposes as *social values*. This meant that the explanations could be augmented with the values promoted, and so inform the choice of the user.

The explanation for Young would now become:

> Where: Plaintiff had not caught the game (F1), the game area is public (F2b) and Plaintiff makes his livelihood from taking game (F3), Plaintiff should win claim. To promote economic usefulness.

And the rebuttal would be:

> In Young, plaintiff and defendant compete (F4). This was not so in Keeble. Thus economic usefulness is not promoted in Young.
>
> Where: the game had not caught the game (F1) and the game area is public (F2b), defendant should win claim. To promote clarity of the law.

4.4. Theory construction

Some researchers have argued that reasoning with legal cases should be seen as a process of theory construction, following the ideas of McCarty [87]. The idea is to construct a theory which will explain the past cases and determine an outcome for the current case. The explanation can then be given in terms of the theory, and competing theories can be evaluated using criteria such as coverage of the past cases, and simplicity, with the simpler theory preferred [51]. One method for theory construction is given in [102], in which each precedent is modelled as a pair of competing rules, one for the plaintiff and one for the defendant and a third rule expressing a priority between these rules according to the outcome of the precedent. This enables the explanation to include the preferences between rules, and the case or cases which established the preference. Construction of theories in which rule preferences are explained by preferences between social values were described in [32]. Construction of these value based theories using heuristic search was implemented in [50]. An alternative approach to theory construction using interactive dialogues can be found in [70] and [128].

5. Argumentation based explanation

All of the above explanations can be seen as *arguments*, reasons for adopting the conclusion. This is natural enough since a legal trial comprises both sides presenting their arguments. This being so it was sensible to look at ideas about argumentation from Informal Logic. This led to the notion of argumentation schemes, first that of Toulmin [123], and later the schemes proposed by Walton [131]. Also in the mid-90s the notion of abstract argumentation [55] emerged, and this too had an important influence on AI and Law [22]. An additional influence was Pollock [95], particularly in identifying different types of attack. A further development is the exploration of structured arguments to be evaluated using abstract frameworks. Ways of representing the structure include ASPIC+ [97] and Carneades [62]. We will discuss explanations based on the schemes of Toulmin and Walton and abstract argumentation frameworks in this section, and ASPIC+ and Carneades in section 6.2, although it should be noted that both make considerable use of argument schemes in generating their arguments.

5.1. Toulmin

One idea to improve explanation from legal systems was to provide an argument structured according to the argumentation scheme of Stephen Toulmin [123]. Independent proposals to use this scheme can be found in [83] and [85]. The idea in all three cases was that presenting the arguments using this structure would assist non-logicians, such as lawyers and jurors, to understand the argument. Toulmin's structure is shown in Fig. 2.

Toulmin's scheme recognises the different roles of statements in an argument:

- Claim: The conclusion of the argument;
- Qualifier: The strength of the claim (certainly, probably, possibly, etc.);
- Data: The premises of the argument;
- Warrant: The inference rule allowing the claim to be inferred from the data;
- Backing: The source of the warrant (in law: statute, case, commentary etc.);
- Rebuttal: A reason why the claim might be though false, or the warrant inapplicable.

Important advantages of the scheme in the legal context are that it incorporates the authority for the warrant, and that it recognises the defeasible nature of legal reasoning by including a rebuttal component which, if true, will block the conclusion. The rebuttal also supports contrastive explanations.

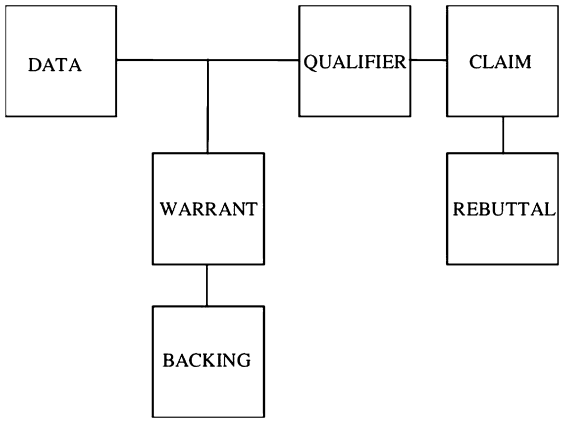

Fig. 2. Toulmin's argumentation scheme.

A computable version of the scheme was provided in [29] which executed an annotated logic program to generate a set of relations instantiating the reasoning as Toulmin argument schemes (e.g. [claim, arg1, findFrorPlaintiff], [data, arg1, private], [backing, arg1, Keeble], etc.). The annotation also excluded certain obvious tests (e.g. 75 > 60) from the explanation, supporting an element of selectivity. Given a suitably annotated version of the rules from section 4, this would give the explanation of *Young* as:

> The argument for the plaintiff is that he captured the animal because he had control and was pursuing his livelihood. This is following *Keeble*. However, this can be rebutted since the defendant was in competition with the plaintiff.
>
> There is no other argument for the Plaintiff.

The explanation presents first the data, then the backing and finally the rebuttal. Warrants were generally omitted, since that the claim followed from the data was considered implicit and so not worth mentioning. Thus the explanation exhibits a degree of selectivity. A more sophisticated presentation of explanations based on this approach was given in [34]. A method for exploring the structure through a dialogue was given in [26], which will be discussed along with other dialogical methods below.

5.2. Other schemes

Throughout the 90s, Toulmin's scheme was the only one prominently used in AI and Law. Around the turn of the century, however, the idea, derived from Walton [131], of using a variety of schemes to represent different kinds of argument (such as Argument from Expert Opinion, Argument from Negative Consequences, Argument from Rules, etc.) was introduced into AI and Law [127]. Argumentation schemes can be seen as a generalisation of the rules of inference. Walton's insight, stemming from his work on fallacies, was to see that certain rules of inference which are, in general, fallacious, may be able to ground presumptively valid inferences provided they are able to satisfy a number of *critical questions*. Thus although given both $P \rightarrow Q$ and Q it is strictly fallacious to infer P, this inference could be presumptively acceptable if no other reason for Q can be shown.

The use of argumentation schemes in law was discussed in [64], where the authors identified five schemes for legal reasoning:

- from position to know
- from ontology
- from rules
- from cases
- from testimonial evidence

Other work used a particular scheme to enable value based argumentation [67], or several schemes ([137] and [103]) to articulate the reasoning of HYPO and CATO. Another approach was that of Grabmair [66], who used a number of argument schemes to express his value judgement formalism for representing legal argumentation.

For our example we will show how the use of schemes can better capture the reasoning in reaching a decision, explaining not only in terms of previous cases, but in terms of the rationales for those decisions. Thus we may argue for the defendant in *Young* using an *argument from authority* (a specialisation of *position to know* to ground an *argument from rule* (R1 in this instance):

Justinian is an authority in this area of law. He said that capture required bodily possession. The plaintiff did not have bodily possession of the fish, and so has not established ownership through capture

One of the critical questions characteristic of the *argument from authority* is whether other authorities disagree. Thus the plaintiff can argue:

Barbeyrac is an authority of this area of law. He denied that bodily possession was necessary to constitute capture.

The defendant can now produce an *argument from case* to establish the preference:

Justinian and Barbeyrac were considered in *Pierson v Post*. The defendant won, showing Justinian was preferred. Therefore the Defendant should win in *Young*.

Several complete reconstructions of the wild animal cases using argumentation are given in a special issue of *Artificial Intelligence and Law*: [20], [98] and [65].

5.3. Abstract argumentation

As well as the opportunities for structured argumentation offered by argumentation schemes, during the late 90s the notion of *abstract argumentation* became increasingly popular. Abstract argumentation derives from the work of Dung [55] and was introduced to AI and Law in [96].

The key notion in [55] is that of an *Argumentation Framework* (AF). An argumentation Framework comprises a pair $< X, R >$, where X is a set of arguments, and R is a set of attack relations between them. In [55] attacks always succeed, so that an attacked argument is acceptable only if none of its attackers are accepted. From this it is possible to identify subsets S of X such that every argument attacking a member of S is attacked by a member of S. If S is also conflict free (no member S is attacked by a member of S), then S is said to be *admissible*. That is, S represents a consistent position, a set of arguments which can be consistently held and which can counter all objections to their members. If an admissible set it maximal it is said to be a *preferred extension*.[4] Properties of preferred extensions include that there is always at least one preferred extension (possibly the empty set), but that there may be several preferred extensions. For example, if the AF comprises just two mutually attacking arguments, each will form a preferred extension: either can be accepted, but not both. Multiple preferred extensions arise when the AF contains one or more cycles of even length ([18], Theorem 2.6).

If we now represent a legal dispute as an AF, then if there is a single preferred extension, then there will be a single clear winner. If, however, there are multiple preferred extensions, then different positions are tenable: this will typically be the case in a legal dispute. Disagreements as to facts always yield two-cycles, but even when the facts are agreed, there can be disagreement on interpretation and points of law. In [17], the wild animals cases (our three example cases, plus several additional cases) were modelled as an argumentation framework, as shown in Fig. 3.

The contents of each argument are not important here, but the plaintiff wins if and only if argument A is in the preferred extension. In fact there are multiple preferred extensions, some with A and some without, arising from the presence of two even length cycles. The two-cycle M-O concerns whether or not Justinian provides an authority that should be followed, and is capable of different resolutions in different jurisdictions or at different times. In the actual series of cases, *Pierson v Post* decided that Justinian should be followed, although the contrary was argued in the minority opinion. The other important cycle, T-S-E-B concerns an allegation of unfair competition (argument T) which arose in *Young v Hitchens*. In practice, the cycle was broken by the court deciding that it could not rule on what constituted unfair competition (argument U), but otherwise there would have been a dilemma: if we accept that the competition was unfair (T) we will also accept E (that Young had done enough to establish possession) and Young will win. Alternatively we can accept the other two arguments in the cycle and Hitchens will win.

The explanation afforded by these systems is of the disagreement between the plaintiff and defendant. The AF is able to identify the cycles and thus present the source of the disagreement, and the consequences of adopting the different positions.

Although this does explain the source of disagreement, it does not explain why the disagreement was settled one way rather than another. In order to address this *Value Based Argumentation Frameworks* (VAF) [18] were used. This approach was used in [19] and [25]. Here the arguments are associated with values (as discussed in 4.3). Now the choice of preferred extension can be explained in terms of value preferences. The AF for *Young* from [19] is shown in Fig. 4.

The framework adopts some of the arguments from *Keeble* interpreted as allowing that Keeble owned the ducks through capture, since his activity was economically useful and he had control of the animals, but adds for *Young* arguments I, J and K. In practice argument K settles the matter (given that the role of the court is more important than considerations of economic usefulness), but in its absence, the plaintiff would win since argument A is defeated whatever the value order

[4] There are many different acceptability semantics for AFs. Three, grounded, preferred and stable, are given in [55], but in subsequent years, many more have been proposed. See [14] for a survey.

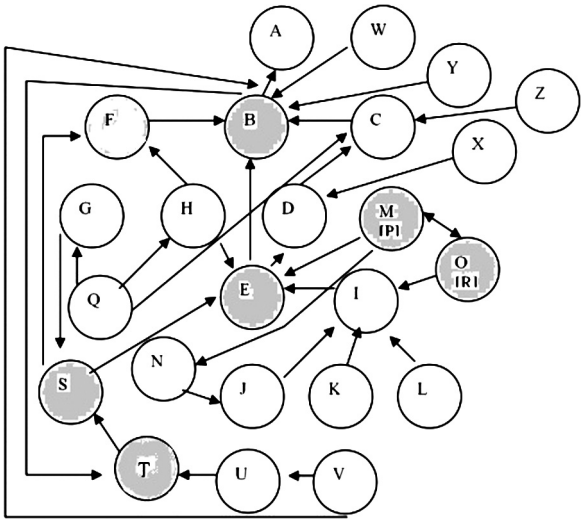

Fig. 3. Argumentation Framework for wild animals cases from [17].

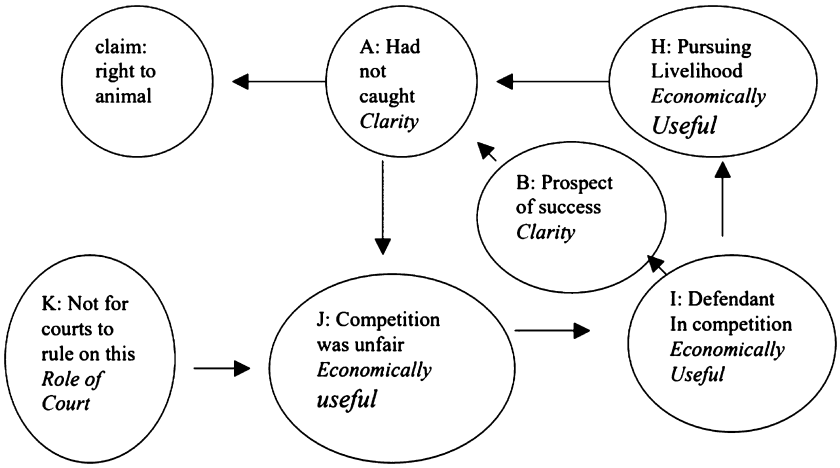

Fig. 4. Value Based Argumentation Framework for *Young* from [19].

(if *clarity* is preferred to *economically useful*, then *B* defeats *A*, otherwise *H* defeats *A*). Thus using VAFs not only identifies disagreements, but can also explain the decisions in terms of value preferences.

5.4. Using Abstract Dialectical Frameworks

Abstract Dialectical Frameworks (ADFs) [44] are a generalisation of abstract argumentation frameworks. ADFs are formed by a three tuple: a set of nodes, a set of directed links joining pairs of nodes (a parent node and its child nodes), and a set of acceptance conditions, expressed in terms of the children. The links show which nodes are used to determine the acceptability (or otherwise) of any particular node, so that the acceptability of a parent node is determined solely by its children.

ADFs have been applied in law to model factor-based reasoning in a number of domains [2] [4]. Used in the legal context, the ADFs' nodes represent statements, which relate to the issues, intermediate factors and base level factors found in CATO's factor hierarchies. The acceptance conditions provide a set of individually sufficient and jointly necessary conditions for the parent node to be accepted or rejected. For leaf nodes, acceptance and rejection is determined by the user, on the basis of the facts of the particular legal case being considered. Collectively, the acceptance conditions can be seen as a knowledge base and they are a feature that provides the modularisation, which is important for being able to easily modify and update the domain knowledge captured in an ADF as the law evolves [1]. Furthermore, the acceptance conditions are used to generate arguments and the ADF structure guides their deployment.

In [3] a methodology for capturing case law (ANGELIC) was presented and it was shown how the CATO trade secrets cases, the automobile exception the 4th amendment, and the wild animals cases discussed previously can be represented as ADFs. Once defined for a domain, an ADF can easily be transformed into a logic program that, when instantiated with the

facts of a case, can determine outcome for the case and the acceptable arguments leading to this decision. The programs reported in [3] demonstrated a high degree of success in replicating the outcomes from the cases used in the experiments, yielding a success rate of over 96% accuracy.

Furthermore, the programs provide output that is highly transparent since they identify precisely the path of reasoning followed through the ADF hierarchy to reach a conclusion on an issue. Below is the output for the case of *Young*, as taken from [3]:

```
?- go(young).
the plaintiff had not captured the quarry
the plaintiff did not own the quarry
plaintiff has good motive
defendant has good motive
plaintiff did not own the land
plaintiff had a right to pursue the quarry
defendant committed no antisocial acts
defendant committed no trespass
no illegal act was committed
do not find for the plaintiff
find for the defendant young
    [rtToPursue,dMotive,pMotive,nc,hp,imp,pliv,dliv]
```

The list of labels given in the final line of the output above are internal names for nodes accepted in the ADF representation of the *Young* case.

In [3] it was also discussed how the programs' output could be built into a more human-oriented explanation. To do this, some re-ordering of the nodes examined was required, along with the addition of some linking text and customisation to refer to issues and base level factors used to invoke a particular clause (as was done in [34]). This yields the following as a sample of what such an explanation could look like for our running example *Young*; clauses from the program output are given in boldface, possible text for issues and base level factors are in italics and linking text is in ordinary font.

> We **find for the defendant. The plaintiff did not own the quarry**. *The plaintiff had not achieved ownership through capture* because **the plaintiff had not captured the quarry** and although *the plaintiff was in hot pursuit* and **plaintiff has good motive, defendant has good motive** also. *Plaintiff did not acquire ownership through ownership of the land* because **Plaintiff did not own the land.** *Plaintiff did not achieve ownership through violated right to pursue* because although **plaintiff had a right to pursue the quarry, defendant committed no antisocial acts, defendant committed no trespass** and **no illegal act was committed**.

The representation in [3] is somewhat more detailed than our earlier examples and includes a third issue (violated right to pursue), which arose in a later case but did not feature in *Pierson*, *Keeble* or *Young*.

To move the work described above into real world applications, a feasibility study was conducted in collaboration with a large law firm to build a practical system using ANGELIC [4]. A body of case law relevant for the business, claims for noise induced hearing loss against employers, was captured as an ADF using the ANGELIC methodology. In this study, identification of usable arguments was crucial to guide case handlers in assessing the strength of a claim and whether or not it had reasonable prospect of defence. The use of ADFs in this task was demonstrated to be highly effective in modelling the domain and assisting case handlers in identifying the arguments relevant for deciding the cases. More recently, this body of work has been extended to examine how ANGELIC can be used to handle reasoning about factors with magnitude [11], as well as Boolean factors, and we see this as a promising area for future development of practical decision-support tools with transparent explanation features.

6. Interactive explanation

The desire to include selective and social elements led to interest in the use of interactive explanations through dialogues. Another attempt to improve the presentation of explanations was through the use of visualisations.

6.1. Dialogue games

During the 90s dialogue games became very popular in AI and Law, e.g. [61], [71], [81], [101] and [16]. A robust implemented system was described in [128]. In the main these systems were designed to allow an adversarial discussion between the two parties, one represented by the computer and one by the user. Social aspects were of particular importance here: there was a recognition that a properly conducted legal case must follow a certain procedure. The modelling of one such procedure was the main motivation of [61]. Such systems did not really explain the reasoning: the users were either expected to come to an understanding of the situation by seeing their arguments met, as in e.g. [101] and [10], or to interactively construct the winning argument which would thus explain the outcome, e.g. [128].

There were, however, approaches which provided not an adversarial persuasion dialogue, but an explanatory dialogue, intended to explain the results of a previous computation. One such dialogue was used in [26] to explain the result established by the production of an argument based on Toulmin's scheme. In this system the user was presented with the claim, and then could use questions to navigate the structure by asking questions appropriate to the current location. For example, when presented with the claim, one could ask for the data (*why?*), the warrant (*because?*), or the possible rebuttal (*unless?*). From the warrant, one could move to the data (*in this case?*), class (*presupposing?*), or backing (*on account of?*).

So the explanation of the Toulmin based argument in 5.1. would yield the following dialogue (with C as the computer and U the user):

[C] The plaintiff has captured the animal
[U] Why?
[C] The plaintiff had control and was pursuing his livelihood
[U] So?
[C] If the plaintiff had control and was pursuing his livelihood he is deemed to have captured the animal
[U] On account of?
[C] The rule was established in *Keeble*
[U] Unless?
[C] The defendant was in competition with the plaintiff

This mode of explanation supports selectivity in particular. Users need only request the elements of which they are unaware. For example, a user familiar with *Keeble* will see that the facts of the current case match the precedent when presented with the data, and so will move straight to the rebuttal.

6.2. Visualisation

We have seen in section 5.3 how diagrammatic representations can support explanation in abstract argumentation frameworks, but diagrams are also widely used for structured explanation.

Although the diagrammatic presentations of Toulmin's argumentation scheme of the sort shown in Fig. 2 were exploited in the early work on Toulmin in [83] and [85], and an alternative proposal for visualisation was given in [82], visualising arguments was greatly popularised by the development of a general purpose tool for argument diagramming, ARAUCARIA [104]. As well as its own standard format, ARAUCARIA supported Toulmin's scheme and also that of Wigmore, who had developed a diagrammatic notation for legal cases [135] used in e.g. [125].

A number of visualisations for AI and Law have been developed. The dialogue in [128] was presented visually, and diagrams based on ASPIC+ [97] have been used in a variety of contexts, e.g. [103] and [99]. Perhaps the leading example of software intended for the visualisation of legal arguments is, however, Carneades [62]. This system presents arguments as a tree which layers claims, their arguments and the premises of these arguments. These diagrams become quite large for complex problems, but the graph taken from [63] relating to why Post did not have possession of the fox in *Pierson v Post* is shown in Fig. 5. Carneades is a highly sophisticated system which draws heavily on the notion of argumentation schemes and currently makes 106 pre-programmed schemes available to ts users [132].[5] The system works by instantiating these schemes from a knowledge base containing facts relating to the case.

In the example in Fig. 5, the claim of argument $a2$ is that Post did not have possession of the fox. Argument $a2$ is an *argument from rule* (see section 5.2 for the argument schemes used by Gordon and Walton). Effectively the rule used is R3 from section 4. The rule itself is justified by three independent arguments from authority, all of which claim that pursuit alone is not sufficient to establish possession.

7. Explaining machine learning

Until very recently the use of machine learning in AI and Law to make and predict decisions was limited, primarily because the explanation facilities were unsatisfactory. The prevalent view was similar to that recently expressed by Robbins:

> "the explanations given by explicable AI are only fruitful if we already know which considerations are acceptable for the decision at hand. If we already have these considerations, then there is no need to use contemporary AI algorithms because standard automation would be available. In other words, a principle of explicability for AI makes the use of AI redundant.... The real object in need of the property of 'requiring explicability' is the result of the process—not the process itself.... Knowing that a specific decision requires an explanation (e.g. declining a loan application) gives us good reason not to use opaque AI (e.g. machine learning) for that decision. Any decision requiring an explanation should not be made by machine learning (ML) algorithms. Automation is still an option; however, this should be restricted to the old-fashioned kind of automation whereby the considerations are hard-coded into the algorithm." [110].

[5] CARNEADES is publicly available at https://github.com/carneades. Last accessed 22nd July, 2020.

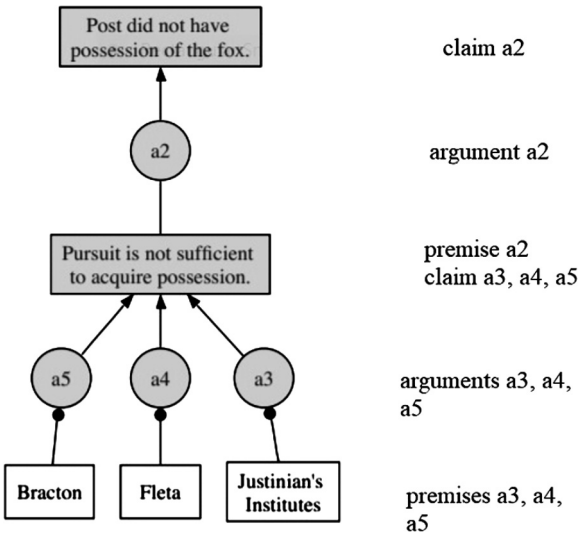

Fig. 5. Carneades explanation of why Post did not have possession, using three Arguments from Authority to ground an argument from rule taken from [63] and annotated to show the various layers.

Essentially the argument was that in order to provide a satisfactory explanation that a domain user could understand, it would be necessary to do the sort of analysis required to build a case or rule based system, so as to identify the terms to use in the explanation. Without such analysis, the explanation would not make sense to the user. As an example, consider [5] a machine learning based prediction system for cases in the European Court of Human Rights which offered by way of explanation "the 20 most frequent words, listed in order of their SVM [Support Vector Machine] weight". One such list, for topic 23 of article 6 predicting violation, is:

> court, applicant, article, judgment, case, law, proceeding, application, government, convention, time, article convention, January, human, lodged, domestic, February, September, relevant, represented

The terms do not look to provide a readily acceptable explanation: some like "court" and "law" one would see as likely to be present in any decision, while others, such as the names of months would appear to be artefacts of the dataset. These are not the terms that a conventional analysis would be likely to identify. But if the analysis to perform a satisfactory explanation has to be undertaken anyway, the construction of a standard case based or rule based system would be the most sensible way to use it.

The analysis remains, however, a substantial, and often daunting task. Therefore, researchers working on AI and Law did explore the use of machine learning, but always recognising that it would be necessary to justify the predictions in terms that could be understood by lawyers and laypeople. Because of the importance of explanation some have proposed hybrid systems, to exploit the strengths offered by machine learning for prediction and explicit representation for explanation [41], [42].

7.1. Past explanations of machine learning

One way in which machine learning techniques were used was to discover rules. In the early 90s neural networks enjoyed a significant amount of popularity, and appeared to be able to produce a very high quality performance on problems which were not well understood. The use of neural networks in law was investigated in [15] and [68]. In [15], to avoid the need for careful analysis to identify relevant features, the dataset used 64 attributes of which only 12 were relevant, to investigate whether the system could learn in the face of this irrelevant information. It was demonstrated that excellent results could be achieved on a legal problem. On a random training set, a success rate of around 98% was achieved. From the trained net, however, investigation revealed that only four of the six conditions that were required to be satisfied were considered, and the high degree of success resulted from multiple conditions being failed. Tested on data which failed only one condition, the success rate fell to around 75%. This paper showed, therefore that a successful model did not necessarily mean that there was a good understanding of the domain which could be used to explain the predictions. If, however, the net could be trained on a dataset which contained passing cases and failing cases that failed on exactly one of the six conditions, good performance could be achieved and all six conditions identified to some extent. Selecting such a dataset would, however, require an understanding of the domain that would be sufficient to permit more traditional techniques to be used.

Other machine learning approaches attempted to extract rules which could be executed using either a standard rule based system or an argumentation based system. Techniques for rule discovery included inductive logic programming (e.g. [91]) and data mining for association rules (e.g. [133]). Both of these papers used the same dataset as [15]. In [91] the

set of rules extracted were compared with the six "ideal" rules. The experiment showed that using CN2 [52] a high level of performance (99%) could be achieved with a defective set of rules. As with the neural net experiment four conditions were correctly identified, also one was partially identified and one was entirely wrongly identified. Using a version of CN2 augmented by argumentation based on an expert's explanation of misclassified data (ABCN2 [91]), the rules could be refined, ending with four correct, one still partial, but far more complete, and one using the right features, but with a threshold of 735 rather than 750. This suggests that in a situation where the rules are unknown, a reasonable approximation could be achieved using ABCN2. These rules could then be deployed in a standard rule based system, and explain the reasoning using the usual facilities. Note, however, this does require the participation of an expert with a good understanding of the domain.

In [133] association rules were refined through a dialogue with moves based on case based reasoning systems such as CATO [6] (including cite, distinguish, counter example, and unwanted consequences of a rule). During the course of the dialogue the rule would be refined so that when the dialogue was complete the winning rule was available to justify and explain the outcome.

Another use of machine learning was the SMILE system [8]. The purpose of SMILE was to enable a pipeline from a natural language description of a case to an outcome. SMILE was based on the domain analysis of CATO [6], which had identified the factors as relevant to the outcome. CATO had identified 26 of these factors, and SMILE identified a separate classifier for each of these factors using a combination of shallow parsing, information extraction, and machine learning techniques. Now given a textual description of a case (a "squib"[6]) SMILE could say whether each of these factors was present or absent, and so provide the cases in the form required by IBP (described in section 4.2 above). The outcome could then be predicted using the IBP system. The machine learning aspect here was not entirely successful: while IBP can achieve better than 90% accuracy on manually ascribed factors, this falls to around 70% when the factors are assigned using SMILE.

What these early experiments showed was that while it was often possible to achieve a good level of performance using machine learning techniques, the rationales for the predictions (and hence the explanations) were often unsatisfactory. Although the quality of the rationales could be improved by expert intervention (selecting the most informative training cases in [15] and explaining misclassified cases in [91]), the effort involved was not dissimilar to that required to analyse the domain for building a knowledge based system.

In all these approaches machine learning was given an intermediate role. Since the system used by the end user only used a product of the machine learning, explanation was not an issue for the end user, although the deficiencies in the knowledge captured by the system definitely was. It should also be noted that all the experiments referred to above, the size of the dataset was, by today's standards, rather small, using hundreds rather than thousands of cases. Currently the use of machine learning in AI and Law is enjoying a significant revival profiting from the widespread availability of large sets of cases and improvements in machine learning techniques. Examples are: [5], [47], and [88]. We will now look at explainability for such systems.

8. Future directions

In this article, we have so far discussed various ways that have been used in existing AI-based legal systems to provide explanations at various levels. Compared to the earlier systems that used association rules and/or a limited set of human-engineered features, modern-day machine-learnt AI and Law systems automatically derive salient features from massive data collections in natural language using deep learning techniques and so pose a complex set of challenges with regard to explainability, which we will discuss in this section.

8.1. Deep learning

Deep learning [59] is a collection of representation learning methods that can automatically learn salient features for a particular task from a given data collection. In classical supervised machine learning algorithms such as support vector machines [126], a human domain expert must first manually specify salient features for a given task, and the learning algorithm will come up with an appropriately weighted and possibly nonlinear combination of those features that can make accurate predictions. The weight associated with a feature can be used as a proxy for determining the importance of that feature for making predictions [93]. This first step of manual specification of salient features is known as *feature engineering* and is often a bottleneck due to multiple reasons such as the cost or unavailability of domain experts and the limited coverage of pre-defined features. Consequently, deep learning methods have gained popularity because they obviate the need for manual feature engineering. Moreover, state-of-the-art performance on a broad range of classification/recognition tasks has been achieved such as in image classification [139], machine translation [56], textual entailment [80] and relation extraction [13] using deep learning methods. Within the legal domain, as well as the outcome prediction systems mentioned above, deep learning-based systems have been proposed for predicting the length of prison sentences [48], detecting medical negligence [37] and extracting information from handwritten documents [122].

[6] A squib is a very brief rendition of a single case or a single point of law from a case in a legal casebook.

8.2. Challenges

Although the ability to automatically learn useful representations for a given set of inputs without any human intervention is a strong advantage of the deep learning-based legal AI systems, it also brings several significant challenges in terms of explainability. Because the features are no longer manually specified, it is difficult to know what features are used by the deep learnt model for making predictions. For example, automatically learnt word representations using deep learning approaches such as word2vec [89] and Global Vector Prediction (GloVe) [94] have shown to encode unfair discriminative gender and racial biases [142,39]. For example, the pre-trained word embeddings predict *homemaker* for the verbal analogy, *man* is to *programmer* vs. *women* is to *?*, which maximises the relational similarity between the two word-pairs: (*man*, *programmer*), (*woman*, *homemaker*). It is common practice to use such pre-trained word embeddings to represent input texts in natural language processing (NLP) applications to improve performance. However, doing so makes those NLP systems biased with regard to legally protected attributes such as gender [112]. Although there have been recent attempts to debias pre-trained word embeddings [75], it has been reported that not all biases are accurately removed by the existing methods [58].

The implication of this problem is particularly worrying in the legal domain because legal decisions must be devoid of any such discriminative biases. Although it is acceptable, for example, to classify sentiment or detect human faces using any available feature in the training data, when it comes to legal decision making we must ensure that the features used by a machine learning-based system are based on existing laws and principles of natural justice. Because deep learning-based methods learn representations automatically from the training data, there is no guarantee that those features will be based on or related to any laws. Unfortunately historic legal data will always be suspect, because bias has been found in a number of instances [47]. It is always important that we do not ossify discredited social attitudes that were formerly prevalent.

8.3. Potential solutions

Attention [12] is a widely used technique for providing explanations into decisions made by deep learning-based models. Specifically, attention is a normalised weight that is learnt that *selects* a subset of features conditioned on a given training instance. It has been shown that attention weights provide useful insights into the decisions made by machine learning-based systems in various application areas [140,92,111]. In Evidence Based Medicine, attention has been used to select sequences of texts from scientific papers that provide evidence for a particular medical procedure or a diagnosis [78]. However, it has been shown that attention alone is inadequate as a form of explanation especially when the number of layers in a deep neural network increases and multiple nonlinear activation functions are used after an attention layer [116, 134].

Branting et al. [43], [42] used Hierarchical Attention Networks (HANs) for predicting the outcomes of World Intellectual Property Organisation (WIPO) domain name dispute cases. These cases have only two possible decisions: granting or denying the request to transfer a domain name to the Complainant, and can be considered as a binary classification problem, given a dispute case. HANs were originally proposed for providing explanation for the predictions made by document classifiers. Specifically, HANs use bi-directional gated recurrent units (GRUs) [49] over pre-trained static word embeddings and concatenate the forward and backward hidden states for each token as a contextualised word representation. Next, sentence embeddings are created as the linearly-weighted sum of the word representations obtained in the previous step, where attention scores are used as the weights. Similar to the way sentence embeddings were created using word embeddings, a second bi-directional GRU is applied over sentence embeddings to create the final embedding for the document. Specifically, each sentence embedding is multiplied by an attention weight that indicates its contribution to the overall meaning of the document and then those weighted sentence embeddings are added up. Because of this two-level (word-level and sentence-level) attention, this model is known as a *hierarchical* attention network. However, it has been shown that highlighting salient words in legal documents alone does not significantly help to reduce the time required to make legal decisions [43]. Often one must read relevant prior cases and cite those as justifications for a legal decision. For this purpose, Branting et al. [43] proposed a semi-supervised approach where they automatically annotate WIPO domain name dispute cases with prior cases with decisions, by measuring the semantic similarity between sentences. We identify attention as a potential future research direction for providing explanations into legal decisions made by deep learning systems.

Providing similar past cases as evidence for a legal decision is a commonly used practice in legal prosecution, and was the basis of the approaches discussed in section 3. If a particular decision is made on a similar case in the past, then following the legal precedence, we must be consistent with our decisions for similar future cases. In machine learning terms this can be formulated as a problem of finding similar past cases with the same decision as we have predicted for the case that we are currently considering. Supervised classification algorithms such as the k-nearest neighbour classifiers and case-based reasoning systems are operating on this principle. A recent example using deep neural networks is reported in [124]. However, deep learning-based classifiers are known to be highly sensitive to *adversarial* examples [120], instances that are carefully perturbed with noise, for which contradictory predictions are made by the deep learnt classifier despite their being no difference that can be noticed by the naked human eye. For example, Goodfellow et al. [60] showed that given an image of a panda, correctly classified by a deep neural network with 57% confidence, we can perturb it with noise that is insignificant to the human naked eye to make the network predict a gibbon with 99% confidence. The implication of this for a legal decision making system is worrying to say the least. We can end up making decisions due to legally

irrelevant minor features or find cases that are not at all legally apposite to support our decisions. Because neural networks are making distributed decisions where a large number of weights in the network are multiplied by the features present in a given instance and summed up and a prediction is made depending on whether this weighted-sum exceeds a threshold or not, we can find adversarial examples that are significantly different in the feature space but result in similar weighted-sums. Adversarial examples have been used to increase the robustness of deep learning systems [60] and we believe that future research in this topic will enable us to design more interpretable and robust prediction systems.

Compared to sub-symbolic approaches such as deep neural nets, symbolic methods are easier to interpret and generate explanations in the form of inference chains. Combining the reasoning capabilities of logic-based symbolic systems and prediction capabilities of deep neural networks[7] to develop hybrid systems is a hotly debated on going topic.[8] Deep learning pioneers such as Yoshua Bengio have strongly argued against hybrid systems proposed by cognitive scientist and the author of *Reboot AI* Gary Marcus claiming that future research in deep learning will be able to provide deep neural nets that can perform inference, making symbolic approaches obsolete. However, at least for the short-term, such hybrid approaches are likely to provide explanations to the decisions made by deep learnt legal prediction systems. For example, Mao et al. [84] proposed a Neuro-Symbolic Concept Learner (NS-CL) that learns representations for visual objects and sentences using neural networks and translates the sentences into executable, symbolic programmes. A neuro-symbolic reasoning module executes these programs in the learnt latent representation space. Because the representational space is continuous, it can be used to easily generalise to previously unseen objects, overcoming the knowledge acquisition bottleneck associated with symbolic approaches. NS-CL shows impressive performance on visual question answering and bidirectional image-text retrieval tasks. There are, however, no examples of this technique being applied to legal prediction.

Although the various machine learning techniques have shown promise and will doubtless attract further research within AI and Law, currently none are able to produce explanations of a comparable standard to the knowledge engineered systems we have discussed. Explanation is an essential feature of legal systems intended to predict case outcomes and so it is crucial that this aspect be developed for machine learning systems intended for deployment on such tasks.

9. Concluding remarks

In this paper we have described the various traditional methods for explaining the reasoning of systems in AI and Law. Despite a recent upturn in interest in machine learning methods, such as [5], [47], [43] and [88], doubts remain about the quality of explanation produced by such systems without the guidance of human experts. Therefore traditional methods continue to be pursued in the development of practical systems [4], and methodologies to support such systems continue to be developed [3]. The importance of, indeed the necessity for, explanations in legal systems, means that this issue cannot be ignored, and without confidence in the explanation, justice cannot be seen to be done.

One feature of legal systems that does ease the task of traditional systems is that in law, although there may be many of thousands of cases available, it is usually unnecessary to consider many of these. Very high performing systems have been constructed on the basis of a limited number of cases: to consider some AI and Law classics: HYPO used fewer than 30 cases [7], CATO used 148 [6], IBP used 186 [45], and reasonable theories have been developed for the wild animals domain with only half a dozen (e.g. [3]). Whereas in machine learning authority comes from the number of cases, in law the level of court, status of the judge, quality of the argument and being followed in subsequent cases are what confer authority. Moreover identifying the leading decisions is not difficult. Although transcripts of all judgements may be available only a small percentage[9] are reported and so available for use in subsequent cases. Of these, only those with important legal significance will be used in subsequent cases. Leading cases can therefore be readily identified and will typically be well known to domain experts. Often they are consolidated in handbooks such as [74] which covers UK tort law. This means that a knowledge engineer can focus on cases that are regarded as significant and ignore the vast majority of cases which may, in any case, include examples where the law was imperfectly implied, which has led to bias in some AI and law machine learning systems [47]. This focus means that it is quite possible to develop practical systems for specific areas of law (e.g. [4]).

Nevertheless the use of powerful machine learning techniques does have its attractions in law, and they will continue to have a significant role in specific areas in which, unlike prediction of decisions, explanation is of lesser importance. For example machine learning has shown to be valuable for the tasks involved in e-discovery [53] and contract review, which can be done by commercially available tools such as Kira Systems.[10] Research will continue also on improving the explanation of prediction systems. However, this facet is of such importance in law that it is essential that the explanations they provide are at least as good as those currently available from systems built from expert knowledge of the domain.

[7] This approach was tried in AI and Law using standard neural networks in [141], but this line of research was not further pursued at that time.
[8] https://montrealartificialintelligence.com/aidebate/.
[9] Around 2%. https://ox.libguides.com/c.php?g=422832\&p=2887381. Last accessed 22nd July 2020.
[10] https://kirasystems.com/. Last accessed 22nd July 2020.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Al-Abdulkarim, K. Atkinson, T. Bench-Capon, Accommodating change, Artif. Intell. Law 24 (4) (2016) 409–427.
[2] L. Al-Abdulkarim, K. Atkinson, T. Bench-Capon, Angelic secrets: bridging from factors to facts in US Trade Secrets, in: Proceedings of JURIX 2016, IOS Press, 2016, pp. 113–118.
[3] L. Al-Abdulkarim, K. Atkinson, T. Bench-Capon, A methodology for designing systems to reason with legal cases using ADFs, Artif. Intell. Law 24 (1) (2016) 1–49.
[4] L. Al-Abdulkarim, K. Atkinson, T. Bench-Capon, S. Whittle, R. Williams, C. Wolfenden, Noise induced hearing loss: building an application using the ANGELIC methodology, Argum. Comput. 10 (1) (2019) 5–22.
[5] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, V. Lampos, Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective, PeerJ Comput. Sci. 2 (2016) e93.
[6] V. Aleven, Teaching case-based argumentation through a model and examples, PhD thesis, University of Pittsburgh, 1997.
[7] K.D. Ashley, Modeling Legal Arguments: Reasoning with Cases and Hypotheticals, MIT Press, Cambridge, Mass., 1990.
[8] K.D. Ashley, S. Brüninghaus, Automatically classifying case texts and predicting outcomes, Artif. Intell. Law 17 (2) (2009) 125–165.
[9] K. Atkinson, Introduction to special issue on modelling Popov v. Hayashi, Artif. Intell. Law 20 (1) (2012) 1–14.
[10] K. Atkinson, T. Bench-Capon, P. McBurney, Parmenides: facilitating deliberation in democracies, Artif. Intell. Law 14 (4) (2006) 261–275.
[11] K. Atkinson, T. Bench-Capon, T. Routen, A. Sánchez, S. Whittle, R. Williams, C. Wolfenden, Realising ANGELIC designs using Logiak, in: Proceedings of Jurix 2019, IOS Press, 2019, pp. 151–156.
[12] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proceedings of ICLR, 2015.
[13] L. Baldini Soares, N. FitzGerald, J. Ling, T. Kwiatkowski, Matching the blanks: distributional similarity for relation learning, in: Proc. of ACL, 2019.
[14] P. Baroni, M. Caminada, M. Giacomin, An introduction to argumentation semantics, Knowl. Eng. Rev. 26 (4) (2011) 365–410.
[15] T. Bench-Capon, Neural networks and open texture, in: Proceedings of the 4th International Conference on Artificial Intelligence and Law, ACM, 1993, pp. 292–297.
[16] T. Bench-Capon, Specification and implementation of Toulmin Dialogue Game, in: Proceedings of JURIX 1998, 1998, pp. 5–20.
[17] T. Bench-Capon, Representation of case law as an argumentation framework, in: Proceedings of Jurix 2002, 2002, pp. 103–112.
[18] T. Bench-Capon, Persuasion in practical argument using value-based argumentation frameworks, J. Log. Comput. 13 (3) (2003) 429–448.
[19] T. Bench-Capon, Try to see it my way: modelling persuasion in legal discourse, Artif. Intell. Law 11 (4) (2003) 271–287.
[20] T. Bench-Capon, Representing Popov v Hayashi with dimensions and factors, Artif. Intell. Law 20 (1) (2012) 15–35.
[21] T. Bench-Capon, HYPO's legacy: introduction to the virtual special issue, Artif. Intell. Law 25 (2) (2017) 205–250.
[22] T. Bench-Capon, Before and after Dung: argumentation in AI and Law, Argum. Comput. 11 (1–2) (2020) 221–238.
[23] T. Bench-Capon, K. Atkinson, Dimensions and values for legal CBR, in: Proceedings of JURIX 2017, 2017, pp. 27–32.
[24] T. Bench-Capon, K. Atkinson, Lessons from implementing factors with magnitude, in: Proceedings of JURIX 2018, 2018, pp. 11–20.
[25] T. Bench-Capon, K. Atkinson, A. Chorley, Persuasion and value in legal argument, J. Log. Comput. 15 (6) (2005) 1075–1097.
[26] T. Bench-Capon, F. Coenen, P. Orton, Argument-based explanation of the British Nationality Act as a logic program, Inf. Commun. Technol. Law 2 (1) (1993) 53–66.
[27] T. Bench-Capon, J. Henderson, A dialogical model of case law dynamics, in: Proceedings of JURIX 2019, 2019, pp. 163–168.
[28] T. Bench-Capon, P. Leng, Developing heuristics for the argument based explanation of negation in logic programs, in: Proceedings of the AAAI Workshop on Computational Dialectics, 1994.
[29] T. Bench-Capon, D. Lowes, A. McEnery, Argument-based explanation of logic programs, Knowl.-Based Syst. 4 (3) (1991) 177–183.
[30] T. Bench-Capon, H. Prakken, A case study of hypothetical and value-based reasoning in US Supreme Court cases, in: Proceedings of JURIX 2009, 2009, pp. 11–20.
[31] T. Bench-Capon, E.L. Rissland, Back to the future: Dimensions revisited, in: Proceedings of JURIX 2001, IOS Press, 2001, pp. 41–52.
[32] T. Bench-Capon, G. Sartor, A model of legal reasoning with cases incorporating theories and values, Artif. Intell. 150 (1–2) (2003) 97–143.
[33] T. Bench-Capon, M.J. Sergot, Towards a rule-based representation of open texture in law, in: C. Walter (Ed.), Computer Power and Legal Language, Quorum Books, New York, 1988, pp. 39–61.
[34] T. Bench-Capon, G. Staniford, PLAID: proactive legal assistance, in: Proceedings of the 5th International Conference on Artificial Intelligence and Law, 1995, pp. 81–88.
[35] D.H. Berman, Developer's choice in the legal domain: the Sisyphean journey with DBR or down hill with rules, in: Proceedings of the 3rd International Conference on Artificial intelligence and Law, ACM, 1991, pp. 307–309.
[36] D.H. Berman, C.L. Hafner, Representing teleological structure in case-based legal reasoning: the missing link, in: Proceedings of the 4th International Conference on Artificial Intelligence and Law, 1993, pp. 50–59.
[37] R. Bevan, A. Torrisi, D. Bollegala, F. Coenen, K. Atkinson, Extracting supporting evidence from medical negligence claim texts, in: Proc. of the 4th International Workshop on Knowledge Discovery in Healthcare Data (KDH) at the 28th International Joint Conference on Artificial Intelligence, 2019.
[38] F.J. Bex, Arguments, Stories and Criminal Evidence: A Formal Hybrid Theory, Springer, 2011.
[39] T. Bolukbasi, K.-W. Chang, J.Y. Zou, V. Saligrama, A.T. Kalai, Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, in: Proceedings of NIPS, 2016, pp. 4349–4357.
[40] L.K. Branting, Building explanations from rules and structured cases, Int. J. Man-Mach. Stud. 34 (6) (1991) 797–837.
[41] L.K. Branting, Data-centric and logic-based models for automated legal problem solving, Artif. Intell. Law 25 (1) (2017) 5–27.
[42] L.K. Branting, C. Pfeifer, B. Brown, L. Ferro, J. Aberdeen, B. Weiss, M. Pfaff, B. Liao, Scalable and explainable legal prediction, Artif. Intell. Law (2020) 1–26.
[43] L.K. Branting, B. Weiss, B. Brown, C. Pfeifer, A. Chakraborty, L. Ferro, M. Pfaff, A. Yeh, Semi-supervised methods for explainable legal prediction, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, 2019, pp. 22–31.
[44] G. Brewka, S. Ellmauthaler, H. Strass, J. Wallner, P. Woltran, Abstract dialectical frameworks revisited, in: Proceedings of the Twenty-Third IJCAI, AAAI Press, 2013, pp. 803–809.
[45] S. Brüninghaus, K.D. Ashley, Predicting outcomes of case based legal arguments, in: Proceedings of the 9th ICAIL, ACM, 2003, pp. 233–242.
[46] B. Buchanan, E. Shortliffe, The MYCIN Experiments of the Stanford Heuristic Programming Project, Addison-Wesley, Reading, MA, 1984.
[47] D.L. Chen, Judicial analytics and the great transformation of American law, Artif. Intell. Law 27 (1) (2019) 15–42.
[48] H. Chen, D. Cai, W. Dai, Z. Dai, Y. Ding, Charge-based prison term prediction with deep gating network, in: Proc. of EMNLP-IJCNLP, 2019.

[49] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: Proc. of EMNLP, 2014.
[50] A. Chorley, T. Bench-Capon Agatha, Using heuristic search to automate the construction of case law theories, Artif. Intell. Law 13 (1) (2005) 9–51.
[51] A. Chorley, T. Bench-Capon, An empirical investigation of reasoning with legal cases through theory construction and application, Artif. Intell. Law 13 (3) (2005) 323–371.
[52] P. Clark, T. Niblett, The CN2 induction algorithm, Mach. Learn. 3 (4) (1989) 261–283.
[53] J.G. Conrad, E-discovery revisited: the need for artificial intelligence beyond information retrieval, Artif. Intell. Law 18 (4) (2010) 321–345.
[54] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Schieber, J. Waldo, D. Weinberger, A. Wood, Accountability of AI under the law: the role of explanation, preprint arXiv:1711.01134, 2017.
[55] P.M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artif. Intell. 77 (2) (1995) 321–357.
[56] S. Edunov, M. Ott, M. Auli, D. Grangier, Understanding back-translation at scale, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, ACL, 2018, pp. 489–500.
[57] A.v.d.L. Gardner, An Artificial Intelligence Approach to Legal Reasoning, MIT Press, 1987.
[58] H. Gonen, Y. Goldberg, Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them, in: Proc. of NAACL, 2019.
[59] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[60] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: Proceedings of International Conference on Learning Representations (ICLR), 2015.
[61] T.F. Gordon, The Pleadings Game, Artif. Intell. Law 2 (4) (1993) 239–292.
[62] T.F. Gordon, Introducing the Carneades web application, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, 2013, pp. 243–244.
[63] T.F. Gordon, D. Walton, Pierson vs. Post revisited: a reconstruction using the Carneades argumentation framework, in: Proceedings of COMMA 2006, IOS Press, 2006, pp. 208–219.
[64] T.F. Gordon, D. Walton, Legal reasoning with argumentation schemes, in: Proceedings of the 12th International Conference on Artificial Intelligence and Law, 2009, pp. 137–146.
[65] T.F. Gordon, D. Walton, A Carneades reconstruction of Popov v Hayashi, Artif. Intell. Law 20 (1) (2012) 37–56.
[66] M. Grabmair, Modeling Purposive Legal Argumentation and Case Outcome Prediction using Argument Schemes in the Value Judgment Formalism, PhD thesis, University of Pittsburgh, 2016.
[67] K. Greenwood, T. Bench-Capon, P. McBurney, Towards a computational account of persuasion in law, in: Proceedings of the 9th International Conference on Artificial Intelligence and Law, ACM, 2003, pp. 22–31.
[68] C. Groendijk, M. Tragter, Statistical and neural approaches to smart-money determination, in: Proceedings of Jurix 1885, 1995, pp. 87–94.
[69] J.C. Hage, Reasoning with Rules: An Essay on Legal Reasoning and Its Underlying Logic, Kluwer Academic Publishers, 1997.
[70] J.C. Hage, Dialectical models in Artificial Intelligence and Law, Artif. Intell. Law 8 (2–3) (2000) 137–172.
[71] J.C. Hage, R. Leenes, A.R. Lodder, Hard cases: a procedural approach, Artif. Intell. Law 2 (2) (1993) 113–167.
[72] J. Horty, Reasoning with dimensions and magnitudes, Artif. Intell. Law 27 (3) (2019) 1–37.
[73] P. Johnson, D. Mead, Legislative knowledge base systems for public administration: some practical issues, in: Proceedings of the 3rd International Conference on Artificial Intelligence and Law, ACM, 1991, pp. 108–117.
[74] M.A. Jones, A.M. Dugdale, Clerk and Lindsell on Torts, Sweet & Maxwell, 2018.
[75] M. Kaneko, D. Bollegala, Gender-preserving debiasing for pre-trained word embeddings, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, 2019, pp. 1641–1650.
[76] I. Lakatos, Proofs and Refutations: The Logic of Mathematical Discovery, Cambridge University Press, 1976.
[77] M. Lauritsen, On balance, Artif. Intell. Law 23 (1) (2015) 23–42.
[78] E. Lehman, J. DeYoung, R. Barzilay, B.C. Wallace, Inferring which medical treatments work from reports of clinical trials, in: Proc. of NAACL, 2019, pp. 3705–3717.
[79] E. Levi, An introduction to legal reasoning, Univ. Chic. Law Rev. 15 (3) (1948) 501–574.
[80] X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, in: Proc. of ACL, 2019, pp. 4487–4496.
[81] R.P. Loui, J. Norman, Rationales and argument moves, Artif. Intell. Law 3 (3) (1995) 159–189.
[82] R.P. Loui, J. Norman, J. Altepeter, D. Pinkard, D. Craven, J. Linsday, M. Foltz, Progress on room 5: a testbed for public interactive semi-formal legal argumentation, in: Proceedings of the 6th International Conference on Artificial Intelligence and Law, 1997, pp. 207–214.
[83] L.S. Lutomski, The design of an attorney's statistical consultant, in: Proceedings of the 2nd International Conference on Artificial Intelligence and Law, ACM, 1989, pp. 224–233.
[84] J. Mao, C. Gan, P. Kohli, J.B. Tenenbaum, J. Wu, The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision, in: International Conference on Learning Representations, 2019.
[85] C.C. Marshall, Representing the structure of a legal argument, in: Proceedings of the 2nd International Conference on Artificial Intelligence and Law, ACM, 1989, pp. 121–127.
[86] L.T. McCarty, Reflections on TAXMAN: an experiment in artificial intelligence and legal reasoning, Harvard Law Rev. 90 (1976) 837.
[87] L.T. McCarty, An implementation of Eisner v. Macomber, in: Proceedings of the 5th International Conference on Artificial Intelligence and Law, 1995, pp. 276–286.
[88] M. Medvedeva, M. Vols, M. Wieling, Using machine learning to predict decisions of the European Court of Human Rights, Artif. Intell. Law (2019) 1–30.
[89] T. Mikolov, K. Chen, J. Dean, Efficient estimation of word representation in vector space, in: Proceedings of International Conference on Learning Representations, 2013.
[90] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[91] M. Možina, J. Žabkar, T. Bench-Capon, I. Bratko, Argument based machine learning applied to law, Artif. Intell. Law 13 (1) (2005) 53–73.
[92] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, in: Proc. of NAACL, 2018, pp. 1101–1111.
[93] A.Y. Ng, Feature selection, l1 vs. l2 regularization, and rotational invariance, in: Proceedings of ICML 2004, 2004.
[94] J. Pennington, R. Socher, C.D. Manning, Glove: global vectors for word representation, in: Proc. of EMNLP, 2014, pp. 1532–1543.
[95] J.L. Pollock, Justification and defeat, Artif. Intell. 67 (2) (1994) 377–407.
[96] H. Prakken, From logic to dialectics in legal argument, in: Proceedings of the 5th International Conference on AI and Law, 1995, pp. 165–174.
[97] H. Prakken, An abstract framework for argumentation with structured arguments, Argum. Comput. 1 (2) (2010) 93–124.
[98] H. Prakken, Reconstructing Popov v. Hayashi in a framework for argumentation with structured arguments and Dungean semantics, Artif. Intell. Law 20 (1) (2012) 57–82.

[99] H. Prakken, An argumentation-based analysis of the Simonshaven case, Top. Cogn. Sci. (2019).
[100] H. Prakken, C. Reed, D. Walton, Argumentation schemes and generalisations in reasoning about evidence, in: Proceedings of the 9th International Conference on Artificial Intelligence and Law, ACM, 2003, pp. 32–41.
[101] H. Prakken, G. Sartor, A dialectical model of assessing conflicting arguments in legal reasoning, Artif. Intell. Law (1996) 331–336.
[102] H. Prakken, G. Sartor, Modelling reasoning with precedents in a formal dialogue game, Artif. Intell. Law 6 (3–4) (1998) 231–287.
[103] H. Prakken, A. Wyner, T. Bench-Capon, K. Atkinson, A formalization of argumentation schemes for legal case-based reasoning in ASPIC+, J. Log. Comput. 25 (5) (2015) 1141–1166.
[104] C. Reed, G. Rowe, Araucaria: software for argument analysis, diagramming and representation, Int. J. Artif. Intell. Tools 13 (04) (2004) 961–979.
[105] A. Rigoni, Representing dimensions within the reason model of precedent, Artif. Intell. Law 26 (1) (2018) 1–22.
[106] E.L. Rissland, Examples in legal reasoning: legal hypotheticals, in: Proceedings of the 8th International Joint Conference on Artificial Intelligence, 1983, pp. 90–93.
[107] E.L. Rissland, Examples and learning systems, in: O.G. Selfridge, E.L. Rissland, M.A. Arbib (Eds.), Adaptive Control of Ill-Defined Systems, Springer, 1984, pp. 149–163.
[108] E.L. Rissland, K.D. Ashley, A case-based system for Trade Secrets law, in: Proceedings of the 1st International Conference on AI and Law, ACM, 1987, pp. 60–66.
[109] E.L. Rissland, K.D. Ashley, A note on dimensions and factors, Artif. Intell. Law 10 (1–3) (2002) 65–77.
[110] S. Robbins, A misdirected principle with a catch: explicability for AI, Minds Mach. (2019) 1–20.
[111] T. Rocktäschel, E. Grefenstette, K.M. Hermann, T. Kočiský, P. Blunsom, Reasoning about Entailment with Neural Attention, preprint arXiv:1509.06664, 2015.
[112] M. Sap, D. Card, S. Gabriel, Y. Choi, N.A. Smith, The risk of racial bias in hate speech detection, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL, 2019, pp. 1668–1678.
[113] U.J. Schild, Open-textured law, expert systems and logic programming, PhD thesis, Imperial College London, 1990.
[114] D.A. Schlobohm, D.A. Waterman, Explanation for an expert system that performs estate planning, in: Proceedings of the 1st International Conference on Artificial Intelligence and Law, ACM, 1987, pp. 18–27.
[115] M.J. Sergot, F. Sadri, R.A. Kowalski, F. Kriwaczek, P. Hammond, H. Cory, The British Nationality Act as a logic program, Commun. ACM 29 (5) (1986) 370–386.
[116] S. Serrano, N.A. Smith, Is attention interpretable?, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL, 2019, pp. 2931–2951.
[117] D.M. Sherman, Expert systems and ICAI in tax law: killing two birds with one AI stone, in: Proceedings of the 2nd International Conference on Artificial Intelligence and Law, ACM, 1989, pp. 74–80.
[118] D.B. Skalak, E.L. Rissland, Arguments and cases: an inevitable intertwining, Artif. Intell. Law 1 (1) (1992) 3–44.
[119] R.E. Susskind, The latent damage system: a jurisprudential analysis, in: Proceedings of the 2nd International Conference on Artificial Intelligence and Law, ACM, 1989, pp. 23–32.
[120] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv:1312.6199, 2014.
[121] S.T. Timmer, J.-J.C. Meyer, H. Prakken, S. Renooij, B. Verheij, A structure-guided approach to capturing bayesian reasoning about legal evidence in argumentation, in: Proceedings of the 15th International Conference on Artificial Intelligence and Law, ACM, 2015, pp. 109–118.
[122] A. Torrisi, R. Bevan, D. Bollegala, K. Atkinson, F. Coenon, Combining textual and visual information for typed and handwritten text separation in legal documents, in: Proceedings of Jurix 2019, 2019.
[123] S.E. Toulmin, The Uses of Argument, Cambridge University Press, 1958.
[124] V. Tran, M. Le Nguyen, S. Tojo, K. Satoh, Encoded summarization: summarizing documents into continuous vector space for legal case retrieval, Artif. Intell. Law (2020) 1–27.
[125] C. Unwin, An object model for use in oral and written advocacy, Artif. Intell. Law 16 (4) (2008) 389–402.
[126] V. Vapnik, Statistical Learning Theory, Wiley, Chichester, GB, 1998.
[127] B. Verheij, Legal decision making as dialectical theory construction with argumentation schemes, in: Proceedings of the 8th International Conference of Artificial Intelligence and Law, 2001, pp. 225–226.
[128] B. Verheij, Artificial argument assistants for defeasible argumentation, Artif. Intell. 150 (1–2) (2003) 291–324.
[129] B. Verheij, Artificial Intelligence as law, Artif. Intell. Law 28 (2) (2020) 181–206.
[130] C.S. Vlek, H. Prakken, S. Renooij, B. Verheij, A method for explaining bayesian networks for legal evidence with scenarios, Artif. Intell. Law 24 (3) (2016) 285–324.
[131] D. Walton, Argumentation Schemes for Presumptive Reasoning, Lawrence Erlbaum Associates, 1996.
[132] D. Walton, Using argumentation schemes to find motives and intentions of a rational agent, Argum. Comput. 10 (3) (2019) 233–275.
[133] M. Wardeh, T. Bench-Capon, F. Coenen, Padua: a protocol for argumentation dialogue using association rules, Artif. Intell. Law 17 (3) (2009) 183–215.
[134] S. Wiegreffe, Y. Pinter, Attention is not not explanation, in: Proc. of EMNLP-IJCNLP, 2019, pp. 11–20.
[135] J.H. Wigmore, Problem of proof, Ill. Law Rev. 8 (1913) 77.
[136] D.E. Wolstenholme, Amalgamating regulation- and case-based advice systems through suggested answers, in: Proceedings of the 2nd International Conference on Artificial Intelligence and Law, ACM, 1989, pp. 63–67.
[137] A. Wyner, T. Bench-Capon, Argument schemes for legal case-based reasoning, in: Proceedings of JURIX 2007, 2007, pp. 139–149.
[138] A. Wyner, T. Bench-Capon, K. Atkinson, Arguments, values and baseballs: representation of Popov v. Hayashi, in: Proceedings of JURIX 2007, 2007, pp. 151–160.
[139] Q. Xie, M.-T. Luong, E. Hovy, Q.V. Le, Self-training with noisy student improves ImageNet classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10687–10698.
[140] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 2048–2057, JMLR.org.
[141] J. Zeleznikow, A. Stranieri, The SPLIT-UP system: integrating neural networks and rule-based reasoning in the legal domain, in: Proceedings of the 5th International Conference on Artificial Intelligence and Law, ACM, 1995, pp. 185–194.
[142] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, K.-W. Chang, Gender bias in contextualized word embeddings, in: Proc. of NAACL, 2019, pp. 629–634.