Explainability of Formal Models of Argumentation Applied to Legal Domain

Michał ARASZKIEWICZ[a,1]
[a]*Department of Legal Theory, Jagiellonian University in Kraków, Poland*
Grzegorz J. NALEPA[b]
[b] *Jagiellonian University, AGH University, Kraków, Poland*

Abstract. This paper deals with the problems of explainability of argumentation models applied to legal domains. The systems based on the theory of Argumentation Frameworks may be in principle fruitfully applied as explanation tools for AI systems enhanced with Machine Learning mechanisms. However, Argumentation Frameworks - as a quickly evolving and diversified field – should be tested with regard to their own explainability. In this paper we provide a set of criteria and we outline a testing procedure towards this goal.

Keywords. Argumentation frameworks, explainability, hybrid systems, legal reasoning.

.

1. Introduction

Explainability has recently arose an important feature of AI systems. While black-box models are nothing new in AI, they have been rapidly growing in popularity. This is due to the fact to the new methods of building them based on big-data in many domains. Today however, these models play an important role in some very sensitive applications of AI systems. This includes medical ones, but also and mainly legal ones. In fact, it could be argued that almost any use of AI could eventually have certain legal implications. This concerns in particular the impact on the subjects' privacy [21]. Therefore, while discussing explainability, we should consider not only the use of AI in legal domain as such, but also the legal dimension of the use of AI as such. Besides its technical merits, AI systems have different limitations regarding their societal acceptance. This is mainly due to the limited trust people have in their operation. Legal analysis of requirements of AI systems can contribute to the building of this trust. In this paper, we do not discuss the legal framework of the application of AI systems. Instead, we investigate the possibility of increasing explainability of AI based decisions via computational argumentation. By definition, argumentation is a rational process of posing reasons for and against a given position is order to choose such conclusion that is best justified in the light of available reasons [comprehensive elaboration of the state of the art: 22]. Argumentation frameworks are a fine example of explainable AI techniques.

[1] Corresponding Author. The writing of this paper was supported by the project K/DSC/004874.

However, they might play an even greater role in the future in explanation generation in hybrid AI systems. Such systems combine black-box models (such as artificial neural networks) with additional explanation facilities (e.g. decision trees). This approach is promising, because it may perform a double role: not only contribute to the explanation of the system's decision as regards the merit, but also analyse its legal implications by providing a reasoning in terms of the decision's acceptability with regard to appropriate legal regulation. In specific cases, the two functions may conflate, in particular where an AI system reasons precisely with the scope of right to privacy or right to explanation [4]. The debate on explainability of AI should therefore take place in the domain of AI and Law research, also because of the growing significance of text analytics and machine learning approach in this field [6]. In particular, machine learning mechanisms are developed to predict the outcome of a legal case (construed broadly; a case in this context means not only a case heard before a court/jury, but any task that involves legal reasoning, such as assessment of a given contractual provision etc.). However in legal context we are particularly interested in receiving a sound justification of the solution to a given problem. A result produced by a learning algorithm may be assessed as correct or accurate, but such result may be reached by accident. There is also a greet degree of bias-related risk in legal reasoning systems. Therefore it is necessary to relate the quantitative and the argumentation-based research on legal argumentation to enhance the transparency of the former. The recent contributions to the domain of transparency of recommending systems [29] provide a basis for analogical investigations in the field of law. Argumentation systems are broadly perceived as a natural candidate for building explanation models for AI. However, before argumentation formalisms are applied to explain the results provided by ML algorithms in the field of law, they should themselves be evaluated with regard to their explainability.

The structure of investigations is as follows. In Section 2 we provide a brief review of the current state of art in the field of computational argumentation and the application of its models to the sphere of law. In Section 3 we develop a scheme for assessment of transparency of argumentation-based models, thus proposing an operationalization of the notion of explainability in this context. Section 4 concludes.

2. Formal and Computational Models of Argumentation and their Applications in Legal Domain

Computational models of argumentation have been intensively investigated since early 1990s, however important earlier work was done also earlier, in 1980s in connection with the rapid growth of interest in the topic of nonmonotonic reasoning modelling. Perhaps the dominant paradigm in the field was started by Dung, who argued for a theory of Abstract Argumentation Frameworks (AAFs) [16]. The idea behind this formalism is simple: certain pieces of information (referred to as arguments) are related means of a binary attack relation. This simple set of primitive concepts enabled Dung and scholars developing this approach to define a set of methods (called semantics) that produce certain sets of arguments (called extensions) which represent those subsets of initially given information that are "justified" in the light of the total set of available information. The formal properties of argumentation frameworks, including complexity features, has been intensively investigated. The characteristic feature of this approach is the existence of different semantics that in certain cases reflect clear intuitions (like in case of

grounded or preferred semantics which respectively correspond to the attitude of a sceptical or credulous person). The methodological status of reasoning modelling by means of argumentation frameworks is debatable. In particular, it is claimed that they simulate human ability to solve problems in intelligent manner [13]. We are of the opinion that this feature strongly depends on the resemblance between reasoning structures in given argumentation framework and those found in reasoning expressed in natural language; and the degree of the said resemblance depends both on the modelled problem and the conceptual richness of a given argumentation frameworks. Simple, early AAFs do not meet this criterion for many categories of problems. This is one of the factors that led to numerous developments in the field (preserving abstract character of the notion of argument):

- Introduction of a second type of relation between arguments: the relation of support, thus leading to development of bipolar argumentation frameworks [3, 14], recently, a third type of relation (neutralisation) was proposed to develop tripolar AFs [29];
- Adding the element of values to AAFs thus developing preference-based and value-based argumentation frameworks [7, 8]
- Introducing the elements of acceptance conditions attached to the elements of reasoning (Abstract Dialectical Frameworks) [12];
- Generalizing the approaches referring to the strength of attacks into the notion of Weighted Argumentation Frameworks [17];
- Allowing group attack relations: joint attacks of certain arguments on other arguments [25];
- Developing a system of multi-level attacks: allowing attacks not only on arguments, but also on attack relations [24];
- Including the intuitions concerning gradual acceptability of an argument in the framework [15];
- Formalizing algebraic operators enhancing reasoning with labels on arguments [13].

Apparently, the on-going tendency consists, first, in extending argumentation frameworks to encompass, as parts of defined vocabulary, certain elements that are explicitly present in argumentation expressed in a given language; second, removing constraints on the possible relations between elements and the types of those elements and third, introducing complexity into the method of acceptance of given sets of arguments, taking into account the intuitions following from the real-life reasoning. However, the inherent limitation of AAFs is that they do not enable to account for the structure of arguments and its role in the process of argumentation. This facet is perhaps the most counterintuitive elements of AAFs theory, for argumentation is naturally accounted for posing reasons (premises) to support or attack a given conclusion. These elements are represented in models of structures argumentation such as ASPIC+ [27] or Carneades [18]. The important feature of these formalisms is the possibility to represent different types of attack on an argument, namely:

- undermining: attack on a premise of an argument,
- undercutting: attack on a relation between premise(s) and conclusion of an argument, and
- rebuttal: attack on the conclusion of an argument.

This feature of structured argumentation models makes them undoubtedly more resembling to argumentation expressed in natural language. The two mentioned models

make use of different formalisms, but it was shown that it is possible to translate Carneades into ASPIC+ [23]. However, formal translatability of certain parts of given formalisms does not mean that they are similarly explainable.

As a matter of course, the above examples of argumentation formalisms do not exhaust the catalogue of the existing approaches. We concentrate on the mainstream, Dungean approach and related research, because of its wide acceptance and extensive elaboration in the argumentation community.

So far, no comprehensive, systematic study has been made into testing the different formal models of argumentation on the basis of instances of legal reasoning. The evaluation of the existing approaches is being done in distributed manner: the authors of a particular model discuss its features on the basis of analysed cases. For instance, one can enumerate the following contributions to the state of the art:

- application of Abstract Dialectical Frameworks to represent case law [1]
- application of ASPIC+ to model reasoning with legal cases [11]
- application of value-based argumentation model to case-based reasoning [9]
- a special issue of Artificial Intelligence and Law journal devoted to modeling one case: Popov v. Hayashi, by means of four different formalisms, including structured argumentation framework with Dungean semantics [10, 19, 28, 31].

However, even in the latter case the formalisms were not compared and evaluated with regard to their transparency, with an intention to increase the transparency of machine learning model. This is not necessarily an objection, because the issue of explainability has only recently become a grand topic in AI and the related research in the legal context is on its preliminary stage. However, the existing gap concerning the assessment of explainability of these models should be filled..

3. Towards Systematic Research on Explainability of Argumentation Formalisms Applied to Legal Domain

The notion of explainability is vague and multi-faceted; therefore the typical approach to this problems consists in adoption of certain measurable criteria (both on the side of the user and on the side of the system) that may be verified in experimental research.

It is first necessary to delineate the class of legal problems against which the explainability of the models should be tested. In our opinion, the most important context is that of judicial application of law, not only because it constitutes the most investigated sphere of legal argumentation, but also because taking into account the potential use of argumentation formalism to explain the functioning of machine learning enhanced predictive models.

The judicial application of statutory law consists of five (interrelated) phases: (1) determination of validity of a norm that is potentially applicable to the current state of affairs; (2) legal fact-finding: stablishing the facts of the case in the process of proof; (3) solving the problems of interpretation of a legal rule; (4) subsumption – determining whether current fact situation qualifies as an instance of the rule's condition and (5) determining the legal consequences of the rule's application [30]. In legal practice, these phases are mutually interrelated, but for the sake of model development it is convenient to consider them in separation. The tasks solved on each of these stages are naturally modelled as argumentative problems. Similar point should be made with regard to application of precedents in Anglo-American law. The process of application of case law

may be subdivided into the following stages: (1) initial characterization of the current state of affairs with regard to its legally relevant features; (2) retrieval of precedents that match the characterization of the case at bar; (3) assessing the similarities and dissimilarities between the current case and the retrieved cases; (4) application of distinguishing argumentation and assessment of counterexamples and (5) determining the outcome in the case at bar. In knowledge-based AI and law research these tasks are best captured with arguments based on knowledge representation structures such as scalable dimensions [5] or binary factors [2], recently extended by the concept of magnitudes [20]. In both legal cultures argumentation on each stage may involve reasoning with values.

The reasoning operations on each stage of any model of application of law may be accounted for as:

- resolving a classification problem (e.g. determining whether a given state of affairs falls under the conditions of a rule or is within the scope of a precedent);
- comparing certain object with regard to certain parameters (as in any case of value balancing, or in case of case comparison with the use of factors);
- assigning consequences to the result of classification or comparison (e.g. by mean od deductive reasoning, defeasible reasoning, analogical reasoning etc.).

Therefore, a good explainable model of legal reasoning should be able to provide justification to the user with regard to the following issues:

- whether a certain object is subsumed under a certain category and why (e.g. on the basis of semantic considerations, prior labeling, etc.);
- what scale (metric) is applied to characterize objects that are subject to comparison, what are the values of parameters of each object on this scale, and why;
- how are the consequences assigned to a certain classificatory decision or a result of comparison, and why; it should be noted that each classificatory decision or result of comparison may have its default consequences which eventually may be trumped by other considerations, for instance following from value-based reasoning.

Let us now consider how argumentation formalisms should be investigated and assessed with regard to the realization of explainability of reasoning and results in legal domain. It should be stressed that this assessment may be performed at three levels of generality at least:

- the level of the argumentation framework as a whole; here, in particular, we may investigate whether the basic conceptual scheme of a given system is sufficient to capture the relevant elements of reasoning;
- the level of application of certain argumentation semantics; even if we agree that the basic conceptual scheme is appropriate for modeling legal reasoning, we may discuss whether a given semantics application is appropriate, for instance with regard to correctness of the results;
- the level of modeling of a concrete reasoning; on this level we may for instance investigate whether the elements of reasoning expressed in natural language are properly transposed into the elements of the framework.

The problem of criteria of explainability of AI systems has already attracted broad attention in the community. The discussion of these criteria takes place first and foremost in the context of evaluation of systems based on statistical methods and enhanced with learning mechanisms. From the point of view of those criteria Argumentation

Frameworks are explainable by definition, because they make use of explicit knowledge and reasoning patterns: there is no need to transpose the quantitative reasoning into qualitative argumentation, because we already begin with the latter. However, the current developments in the theory of AFs, enhancing their expressive power, at the same time consists in introduction of more and more complicated logical and mathematical tools, thus decreasing the transparency of elements of knowledge bases and reasoning patterns. Thus it is worthwhile to recall a part of the classic set of requirements that were discussed in earlier AI and law work in connection with representation of rules and exceptions in defeasible logic systems [26], applied accordingly to the problem of modeling legal argumentation:

- structural resemblance: preserving the structure of knowledge units and argumentation with regard to natural language expressions;
- modularity: formalizing parts of the domain without taking into account the whole domain at the same time (practically important for validation and maintenance, but also increasing explainability because of the limited capacities of actual user to handle too much information at one time);
- expressiveness: the formalization should be able to capture all distinctions that are important in natural language reasoning.

We think that the above criteria may be fruitfully applied as criteria of explainability of models of legal reasoning based on Argumentation Frameworks. However, we think that it is necessary to add another important criterion:

- substantial resemblance: the reasons that justify certain conclusions should be identical or at least significantly similar to those accepted by an experienced expert. In other words: not only the structure of reasoning, but also its merit (content) should be in a certain similarity relation between natural language reasoning and reasoning represented in an AF.

The substantial resemblance requirement is more important for explainability that the standard criterion of accuracy of result broadly adopted in ML-enhanced systems. As discussed above, a correct answer may be yielded by a system by accident, or through the flawed, fallacious reasoning, also on the basis of distorted or false data. The explainable system should be ready to answer the why-questions in a matter similar to the expert user.

Taking into account the set of criteria we may outline the process of testing explainability of argumentation models for legal domain.

- the choice of use case (UC);
- formalizing the knowledge elements present in the UC with the tested argumentation formalisms;
- enabling the systems to generate the conclusion;
- comparison of the conclusions generated by systems to the ones adopted by expert users;
- careful investigation of each step of performance of the system with regard to the adopted criteria, with an applied scale (such as Likert scale or another);
- evaluation of results and development of sets of postulates with regard to: (1) applicability of a given Argumentation Framework to legal modeling reasoning; (2) applicability of a given semantics and (3) formalization of a concrete case.

Several *a priori* hypotheses can be made at this point, with a reservation that they may be falsified in the course of experimental work. First, structured argumentation systems should be assessed higher than abstract argumentation frameworks, because legal

reasoning essentially involves the analysis of relation between premises and conclusions of an argument, and not only the analysis of conflicts between arguments. However, it should be noted on the contrary that AAFs enable to map the notion of "argument" to an element of natural language argumentat (and not to a given argument taken as a whole), which may decrease the importance of this drawback. Second, the formalism should enable representation not only of attack relations, but also support relations; this hypothesis favors bipolar over classical argumentation frameworks. Third, taking into account the expressiveness, structural resemblance and substantial resemblance as criteria, and the role of value judgments in legal reasoning on the other hand, the testing should presumably favor the formalisms that used the notion of values explicitly. Fourth, because in legal reasoning we use elements that follow from different sources and that the "pedigree" of elements is an important factor, the argumentation framework should enable some labeling to express this aspect. Fifth, as comparison of objects (in particular weighing of values) involves scalable reasoning, the frameworks that enable weighted relations or gradual attacks will be preferred by default. However, the choice of proper scale and metrics is a complicated issue: too fine-grained scale may decrease explainability.

4. Conclusions

The current developments of AI systems have more and more influence on the life of individuals and societies. The problem of explainability of decisions made with the support of those systems as well as procedures they are based on has become an important social and legal issue. Argumentation Frameworks may in principle be fruitfully used as tools of explanation of the AI systems' operation. However, AFs themselves, as models developed in a complex, diversified and quickly evolving fieldof research, should themselves be tested with regard to their transparency and explainability. In this contribution we have outlined a general procedure for such testing, subject to future development.

5. References

[1] Latifa Al-Abdulkarim, Katie Atkinson, Trevor J. M. Bench-Capon: A methodology for designing systems to reason with legal cases using Abstract Dialectical Frameworks. Artif. Intell. Law 24(1): 1-49 (2016)

[2] Vincent Aleven 1997. Teaching Case-Based Argumentation Through a Model and Examples, PhD Dissertation, Teaching Case-Based Argumentation Intelligent Systems Program, University of Pittsburgh

[3] Leila Amgoud, Claudette Cayrol, Marie-Christine Lagasquie-Schiex 2004. On the bipolarity in argumentation frameworks. NMR 2004: 1-9

[4] Michał Araszkiewicz. 2018. Sztuczna Inteligencja i prawo do wyjaśnienia (Artificial Intelligence and the right to explanation), Trzeci Sektor IV/2018, forthcoming.

[5] Kevin Ashley. Modeling legal argument. Reasoning with Cases and Hypotheticals. MIT Press, Cambridge: Mass., 1990.

[6] Kevin Ashley, 2017. Artificial Intelligence and Legal Analytics. New Tools for the Law Practice in the Digital Age, Cambridge: Cambridge University Press

[7] Trevor J. M. Bench-Capon: Value-based argumentation frameworks. NMR 2002: 443-454

[8] Trevor J. M. Bench-Capon: Persuasion in Practical Argument Using Value-based Argumentation Frameworks. J. Log. Comput. 13(3): 429-448 (2003)
[9] Trevor J. M. Bench-Capon, Katie Atkinson, Alison Chorley: Persuasion and Value in Legal Argument. J. Log. Comput. 15(6): 1075-1097 (2005)
[10] Trevor J. M. Bench-Capon: Representing Popov v Hayashi with dimensions and factors. Artif. Intell. Law 2012 20(1): 15-35
[11] Trevor J. M. Bench-Capon, Henry Prakken, Adam Z. Wyner, Katie Atkinson: Argument schemes for reasoning with legal cases using values. ICAIL 2013: 13-22
[12] Gerhard Brewka, Stefan Woltran: Abstract Dialectical Frameworks. KR 2010
[13] Maximiliano Celmo Budán, Gerardo I. Simari, Ignacio Darío Viglizzo, Guillermo Ricardo Simari: An approach to characterize graded entailment of arguments through a label-based framework. Int. J. Approx. Reasoning 82: 242-269 (2017)
[14] Claudette Cayrol, Marie-Christine Lagasquie-Schiex 2005. On the Acceptability of Arguments in Bipolar Argumentation Frameworks. ECSQARU 2005: 378-389
[15] Claudette Cayrol, Marie-Christine Lagasquie-Schiex: Graduality in Argumentation. J. Artif. Intell. Res. 23: 245-297 (2005)
[16] Phan Minh Dung, 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. Artif. Intell. 77(2): 321-358
[17] Paul E. Dunne, Anthony Hunter, Peter McBurney, Simon Parsons, Michael Wooldridge: Weighted argument systems: Basic definitions, algorithms, and complexity results. Artif. Intell. 175(2): 457-486 (2011)
[18] Thomas F. Gordon, Douglas Walton: The Carneades Argumentation Framework - Using Presumptions and Exceptions to Model Critical Questions. COMMA 2006: 195-207
[19] Thomas F. Gordon, Douglas Walton: A Carneades reconstruction of Popov v Hayashi. Artif. Intell. Law 20(1): 37-56
[20] John F. Horty: Reasoning with dimensions and magnitudes. ICAIL 2017: 109-118
[21] Edwards, Lilian, Veale, Michael. 2017. Slave to the Algorithm? Why a 'Right to an Explanation' is Probably not the Remedy You Were Looking For. Duke Law and Technology Review, 16: 18–84.
[22] Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, Jean H. M. Wagemans. 2014. Handbook of Argumentation Theory. Dordrecht: Springer Science+Business Media.
[23] Bas van Gijzel, Henry Prakken: Relating Carneades with Abstract Argumentation. IJCAI 2011: 1113-1119
[24] Sanjay Modgil, Trevor J. M. Bench-Capon: Metalevel argumentation. J. Log. Comput. 21(6): 959-1003 (2011)
[25] Søren Holbech Nielsen, Simon Parsons: A Generalization of Dung's Abstract Framework for Argumentation: Arguing with Sets of Attacking Arguments. ArgMAS 2006: 54-73
[26] Henry Prakken. 1997. Logical Models for Modelling Legal Argument. A Study of Defeasible Reasoning in Law. Dordrecht: Springer.
[27] Henry Prakken: An abstract framework for argumentation with structured arguments. Argument & Computation 1(2): 93-124 (2010)
[28] Henry Prakken: Reconstructing Popov v. Hayashi in a framework for argumentation with structured arguments and Dungean semantics. Artif. Intell. Law 2012, 20(1): 57-82
[29] Antonio Rago, Oana Cocarascu, Francesca Toni. 2018. Argumentation-Based Recommendations: Fantastic Explanations and How to Find Them. IJCAI 2018: 1949-1955.
[30] Jerzy Wróblewski. 1992. The Judicial Application of Law. Dordrecht: Springer.
[31] Adam Z. Wyner, Rinke Hoekstra: A legal case OWL ontology with an instantiation of Popov v. Hayashi. Artif. Intell. Law 2012, 20(1): 83-107