An Evaluation of Methodologies for Legal Formalization

Tereza Novotná[1][0000−0002−1426−4547] and Tomer Libal[2][0000−0003−3261−0180]

[1] Institute of Law and Technology, Masaryk University, Brno, Czech Republic
tereza.novotna@law.muni.cz
[2] University of Luxembourg, Luxembourg
tomer.libal@uni.lu

Abstract. Legal formalization is a necessary step for automated legal reasoning because it aims to capture the meaning of legal texts in a machine-readable form. Therefore, there are many attempts to create better legal formalization methods and to make legal reasoning techniques more efficient. However, these methods are rarely evaluated and thus, it is difficult to recognize the "good" legal formalization method. In this article, the authors provide a categorization of necessary properties of a "good" formalization that is based on a literature review of recent state-of-the-art methods. They conclude in favour of the legal experts' evaluation method as the most suitable one for assessing the quality of legal formalization.

Keywords: legal formalization · evaluation · literature review

1 Introduction and motivation

The relationship between the law and logic and its characteristics is an old theoretical and philosophical issue. The intuitive need to find logical principles in legal rules goes far beyond the pragmatic (practical) reasons of why the application of logic in law is being examined today. Therefore, an extensive amount of research is dedicated to exploring this relationship, or specifically to find the logical representation of legal rules, to be able to automate them and to automatically reason over them.

This task became even more interesting and also more important with the development of information technologies, with the availability of publicly accessible legal data and mainly with new natural language processing methods or artificial intelligence in general. As the amount of legal text is constantly rising, the need for more advanced, faster and more intuitive tools for automatic legal reasoning is increasing as well. On the other hand, so does the expectations of lawyers and users regarding the user-friendliness and the accuracy of such tools. Therefore, there are many attempts of many research teams to present better methods and tools for legal formalization and automatic legal reasoning. However, the crucial question remains unanswered - how to recognize a good legal formalization method?

A logical solution adopted from software engineering is the evaluation of such a system, tool or method. As stated in [11], the evaluation of a system is necessary for three main reasons: to demonstrate accountability, gain knowledge and enhance development. The evaluation of methods is also highly recommended by the members of the artificial intelligence and law community themselves, as it is a sign of maturity and essential scientific rigour [13] and it supports an empirical assessment of the research efforts in both qualitative and quantitative ways [8]. In this field, different evaluation methodologies are used to fit the different purposes and implementations of a wide range of methods. For example, a widely adopted approach is Context Criteria Contingency-guidelines Framework (CCCF) as proposed in [17] for the evaluation of legal knowledge base systems. For the evaluation of legal ontologies, there is a well established validation methodology as it is described in [29].

However, the evaluation and its methodology is a research problem itself, with its many purposes and many possible approaches. Therefore, there are efforts within the research community to theoretically frame the evaluation methodology as well. Six categories of evaluation methods used in artificial intelligence and law field are defined in [14], and these are:

1. Gold Data: evaluation performed with respect to domain expert judgments (e.g., classification measurements or measures on the accuracy, precision, recall, F-score, etc.);
2. Statistical: evaluation performed with respect to comparison functions (e.g., unsupervised learning: cluster internal-similarity, cosine similarity, etc.);
3. Manual Assessment: performance is measured by humans via inspection, assessment, review of output;
4. Algorithmic: assessment made in terms of performance of a system, such as a multi-agent system;
5. Operational-Usability: assessment of a system's operational characteristics or usability aspects;
6. Other: those systems with distinct forms of evaluation not covered in the categories above (task-based, conversion-based, etc.).

The methodological proposal presented in this article follows the theoretical distinction of different categories of evaluation. However, it respects the specific nature of legal formalization, as it relies heavily on human input in all of its different phases. Given the categorization above, the authors consider manual assessment as the best fit for the evaluation of legal formalization methods and they argue this statement and explain it later in this article.

Based on the literature review of the most recent state-of-the-art methods of legal formalization (in Section 2) and on the review of whether and how these methods are evaluated (in Section 3), the authors here define the necessary properties for legal formalization method to be practically applicable for legal reasoning (in Section 4). They further argue (in Section 5) that human-centered evaluation is the most suitable method to answer the question: *What constitutes a "good" legal formalization method and how to evaluate it?*

2 Literature overview of recent legal formalization approaches

Legal formalization is defined as a translation of the legal (regulatory) text into its logical representation with the preservation of its legal meaning. As such, it is usually the first and necessary step in the process of automatic legal rule-based reasoning. The correct logical representation of legal rules is used as a basis for answering questions about a formalized legal text, the answers can be deduced from the logical representation as logical consequences of formalized legal rules using an inference engine. The important difference between the two phases is the difference in the use of advanced technologies during the process. While in the latter phase, the adoption of artificial intelligence leads to more efficient and faster inference engines, the first phase of legal formalization is still at least partly performed manually based on the know-how of several involved experts (legal expert, programming expert, logician etc.). This situation clearly shows that legal formalization is a complicated task with several related problems further discussed in Section 4. Nevertheless, it still attracts research groups that propose solutions to some of the related issues and methodologies for better legal formalization in general, because of the great potential of such a research direction.

For a broader context, it is necessary to mention that legal formalization is a standard part of some other lines of research. A review of all these approaches is beyond the scope of this paper, but at least a few important examples should be mentioned. First of all, it is legal argumentation, its modelling and automation, both case-based and rule-based. Some recent research includes for example [32], which deals with the formalization of legal cases for the purpose of argumentation and legal reasoning, but which is based on the methodology of the CATO project introduced in [3] (based on [5]). For completeness it is necessary to mention that this research was also evaluated by a group of law students in [2]. Additionally, the connection between formalization of legal cases and rules in order to provide legal argumentation framework is introduced in [38]. A related direction is that of legal interpretation and its logical representation, which is then often applied precisely for the purpose of the aforementioned argumentation, for example in [4] or in [34].

In view of our objective, which is a an overview of recent attempts in legal formalization of legislation, we provide a non-exhaustive overview of the most recent state-of-the-art efforts in this direction.

One of the few proposed solutions with a computer-assisted methodology for extracting logical representation from legal text is in [28]. In this work, the authors propose a detailed methodology of extracting logical rules based on a very precise linguistic analysis of legal regulatory text in the Japanese language. Logical relations in the legal text are stored in the Davidsonian Style. However, the authors limited the interpretation of the legal text on the assignment of different terms to similar meanings without further contextual semantics. The possibility of different interpretations of a legal text (also related to the open

texture characteristics of legal text) is an issue widely discussed in subsequent research and essential characteristics of an applicable legal formalization method.

Another research team Bartolini et al. in [7] used a logical representation of legal texts for finding correlations between two related legislations - GDPR and ISO 27018 standard - to make the personal data protection compliance checking easier. GDPR (and related ISO standard) was selected because of its general importance and wide applicability. At the same time, many following attempts in legal formalization were experimentally applied to the same regulation. The main idea was that logical representations of the provisions from both legislations are language-independent and therefore, they may be helpful when searching for document-to-document correspondences. These correspondences may be subsequently helpful for compliance checking with both legislations. The authors used a methodology based on a knowledge base in the form of legal ontology and Reified Input/Output Logic for creating logical formulae of legal provisions. The interpretation of different legal terms appearing in the text and the logical relationships among the terms is found by a legal expert, although authors argue that because of the XML-based ontology, the interpretation of a legal text can be changed easily along with the ever-changing characteristics of law.

In the following research within the same project, the authors presented a DAPRECO knowledge base in [33]. DAPRECO is based on the PrOnto legal ontology of legal concepts in GDPR (described in detail in [30]) and logical formalization of legal rules is performed in Reified Input/Output Logic. The whole knowledge base is coded in LegalRuleML. The authors declare that this combination allows them to create an overview of legal concepts in GDPR thanks to the ontology and at the same time, to automatically reason over logical rules extracted from the text and related to the ontology concepts. Additionally, using the Reified Input/Output logic allows them to deal with some widely discussed issues related to legal formalization, such as nested obligations and permissions, exceptions and handling more than one interpretation of a legal text.

To discover, whether their methodology is meaningful for actual use for automatic legal reasoning, the authors propose an evaluation methodology of their approach and demonstrate it on a small-scale evaluation experiment with legal experts in [8]. This evaluation experiment and its results are further discussed in Section 3.

Palmirani and Governatori in [30] used LegalRuleML to formalize the GDPR for automatic legal reasoning. Additionally, they combine formalization of legal text with PrOnto legal ontology as a legal concepts base in an integrated framework. Their goal is GDPR compliance checking. They use defeasible logic and legal experts using graphical interface to formalize the legal text. However, they do not deal with multiple interpretations and they do not specify the cooperation with legal experts. Secondly, they don't provide use examples and any evaluation of correctness or usability of the system.

A different approach is described in [27]. In the cited work, the Prolog language is used to capture logical relations in the GDPR and model compliance checking system on this representation. They demonstrate this approach on two

articles of the GDPR. Despite the fact that their work has different focus than high quality legal formalization, it is still a necessary part of it. The authors translate the text of the articles into logical representation and interpret it themselves without any further evaluation. On the other hand, they provide precise description of their interpretation and formalization of the text of the GDPR.

Part of the research related to the formalization of the statute law of the USA (specifically tax law) was proposed in [21, 31, 20]. Firstly, Lawsky in [21] argues for the default logic to be used for the formal representation of statute law and her opinion was based on supporting defeasible reasoning, which is a natural characteristic of law. In [31] the authors suggest a practical application of default logic to tax statute (Internal Revenue Code). They use regular expressions and automatic parsing to automatically translate the legal text into logical representations in default logic. The results were promising, but not accurate enough to be applied as is. The authors don't consider the problem of different interpretations of the legal text, on the other hand, their proposed method is more efficient in terms of time spent on the manual translation of the legal text to logic. Lawsky then suggests in a subsequent article [20] to focus on tax forms, which are easier to automatize, than the legal text itself.

One of the very recent efforts in legal formalization is the CATALA project described in [26]. Its highly ambitious goal is to provide a generally usable logical language for legal formalization based on prioritized default logic. The authors declare that their proposed logical language and its implementation is rich enough to cover all main issues related to the legal formalization. Additionally, they declare that such an implementation is comprehensible enough to be used (or edited) by lawyers with no background in logic. They support this declaration with an evaluation experiment described in detail in Section 4.

Last research direction in this non-exhaustive list is the one originally proposed in [25]. The authors of this paper shifted the focus from creating exhaustive knowledge bases combining legal ontologies and legal formalization to more human centered one. They propose a tool, called the NAI tool, for the formalization of legal text that is ready to be used by lawyers and technical laymen themselves. Their approach aims to provide a user friendly graphical interface, which can be used by lawyers to annotate logical relationships in the text according to their interpretation. This interpretation may be easily rewritten and the tool supports creating multiple interpretations at the same time. The tool allows automatic reasoning over the logical formalization afterwards. The authors propose some use cases to demonstrate the functionality. However, this approach is not further evaluated. On the other hand, the annotation editor is freely available online[3] and its source code is open source.[4] The NAI tool was subsequently used for demonstration of inconsistency checking within a legislation in [24] and it was extended to provide transparent methodology for legal formalization in [23].

As it was showed in this Section, legal formalization and automatic legal reasoning is widely employed within the community and research teams present

[3] https://nai.uni.lu/
[4] https://github.com/normativeai

new methods and approaches to provide a good formalization of legal text. Nevertheless, the evaluation of such efforts is not very common. Only few of cited works evaluated their results. The evaluation methods and results are described in the following Section.

3 Overview of evaluation methods of legal formalization

The importance of evaluation studies following the application of different artificial intelligence methods for either legal information retrieval or automatic legal reasoning is emphasized within the artificial intelligence and law community itself. In [12], the authors state that the evaluation of the experiments and the methods *"expedites the understanding of available methods and so their integration into further research"*. The authors in [14] argue, that *"a performance-based 'ethic' signifies a level of maturity and scientific rigor within a community"*. However, the meta-analysis of research studies in the field of artificial intelligence and law in [18, 13, 14] shows that great part of studies does not contain any kind of evaluation whatsoever (their typology of evaluation is presented in Section 1).

In this article, the authors narrow the scope of the meta-analysis on evaluation in artificial intelligence and law field and focus specifically on evaluation of state-of-the-art methods in the field of legal formalization. Moreover, they stress the importance of the human-centered evaluation methodology in this field. According to the authors, this type of evaluation is usually the most time consuming, and secondly, it needs to be very precisely designed to provide meaningful and objective results. For comparison, in [14], the authors discovered that only 22 percent of evaluated works in artificial intelligence and law field, that they included in their meta-analysis, employed manual assessment as an evaluation methodology. Despite these constraints, the authors believe that human-centered evaluation experiments bring the most meaningful and significant results. This opinion is based on the specific characteristics of legal formalization, which are discussed further in this article. E.g. [6] can be mentioned as an example of good practice, even though the evaluated method differs from the one that is the subject of this review.

The first reason is the high interdisciplinary character of legal formalization. In the majority of cases, it is necessary to employ both computer scientists (and even logicians) and lawyers or domain experts generally. The understanding of legal text, capturing its correct meaning and translating it into logical representation is a difficult task requiring different knowledge. Secondly, this method usually depends strongly on the cooperation with humans (programmers, logicians, legal experts, users) during the whole process of legal formalization and at the same time, humans are expected to use it in practice in the end. Therefore, it is ideally supposed to be evaluated by humans as well. And the third and pragmatic reason is that there are not many other well established evaluation methods which could be applicable - such as statistical measures, comparison to golden dataset etc.

The authors are aware of two recent works adopting a legal experts' evaluation - [8, 26]. Both of the works adopt different evaluation methodologies. In [8], a small-scale evaluation experiment is conducted with a group of four lawyers (two with and two without knowledge of deontic logic). This group of legal experts is presented with a small part of formalized legislation in a human-readable form, and they are asked to answer *yes/no* questions about the formalized legal text in the form of a questionnaire. These questions target the *accuracy, completeness, correctness, consistency* and *conciseness* of such a formalization. Their results are promising, mainly for the human-readable form of formalization, however the significance is not that high given small number of participants. Additionally, lawyers are evaluating logical representation translated into human-readable form, although this form is created only for the purpose of this evaluation experiment and there is no intend of authors expressing its broader use. The interrater agreement is measured as well, with satisfactory results.

In [26], the authors claim that this evaluation experiment is only the initial user study, without a proper scientific methodology. The evaluation group consists of 18 students of law, which is a more significant size of the evaluation group. Also, in this case, students are evaluating the actual code that captures the formalized legal text. They are evaluating several questions related to the code targeting its correctness, completeness, comprehensibility etc. The law students are answering in open answers, i.e as free text and their answers are classified by authors as either positive, negative or mixed. That is a problematic methodology reduces the possibility to compare or generalize the results, as it requires interventions by the authors and their interpretation of evaluation answers. The evaluation methodology is providing explanation of this approach and explanation of the meaning of the code that captures the logical representation of legal text to the students. The authors assess these results as promising despite several methodological issues.

The reviewed evaluation experiments are good first steps towards objective evaluation of legal formalization methods, however, methodological flaws limit both of them in presenting significant proves of the quality of the evaluated methods, as it is further discussed in following sections.

4 What constitutes a "good" formalization?

One of the main conclusions from our overview of the different approaches to legal formalization is that there is no clear answer to this question, nor there is a faithful and objective methodology to recognize it. The different approaches that were presented in previous sections usually target narrowly selected parts of legal text and their application is, at best, evaluated within the framework of a small-scale evaluation experiment over an arbitrarily selected sample of legislation. The evaluation experiment, if any, usually aims to convince the expert community of the accuracy and applicability of the method or approach. But, different evaluation methodologies prevent the results from being generalized

and compared with each other. At the same time, there are ambitious efforts to deal with some well-known and described issues related to legal formalization.

A summary by Branting [10] suggests a typology of the main problems related to the formalization of legal texts, which restricts the general use of any of the cited methods. He classifies the issues into two main categories - *"the challenge of efficiently and verifiably representing legal texts in the form of logical expressions; and the difficulty of evaluating legal predicates from facts expressed in the language of ordinary discourse"*. These two categories are rather general concepts summarizing related issues and often overlap. However, they are a good starting point for the discussion described in this section and we therefore introduce them next.

In the first category, he focuses on a problem related to the correctness of a formal representation of legal rules and refers to it as the issue of *fidelity and authoritativeness* of logical representation. According to the author, the *fidelity* element refers to the actual correctness of a formalization and the *authoritativeness* refers to the binding nature of such a formalization. The problem of having more than one correct interpretation of a legal text is closely linked to the very nature of written law. Nevertheless, the correctness of such an interpretation is highly related to other extralegal sources, such as social situations or time. Furthermore, even there is one widely accepted (or correct) interpretation, the logical language must be too complex to represent correctly and completely such an interpretation while preserving the legal meaning. Moreover, Branting argues that in the case when legal text is correctly interpreted and accurately translated to its logical representation, there is no legal guarantee of the binding nature of such logical representation (the problem of authoritativeness).

In the second category, Branting defines the problem of a gap between legal terms and ordinary language of facts or real-life situations, which is further emphasized with the assignment of logical formulae to legal text. The subsumption of general facts to legal rules may be a cognitively demanding process even for lawyers and legal experts in a field, all the more for laymen. Branting relates this problem to the vagueness, ambiguity or open texture characteristics of legal text, which has been extensively described in related literature [15, 37] and offers some *ad hoc* solutions.

In this paper, we choose an approach that in some respects is based on this typology, but which is supplemented with observations from additional research and the studies referred to above. This additional information is rather recent and presents new important issues, which have not been addressed in Branting's overview [10].

By classifying the important issues related to legal formalization, we define four essential parameters of a good quality formalization: correctness, transparency, comprehensibility and multiple interpretations support. Furthermore, the authors suggest how to evaluate these parameters for the comparison of different approaches.

4.1 Correctness

The *correctness* of the logical representation of a certain legal text and its meaning is indisputably the most important parameter and *conditio sine qua non*, some research directions use the concept of isomorphism as a mapping between a legal rule and its representation [9]. Branting in his typology uses the term *fidelity* of such a representation. Bartolini et al. [8] even distinguish different dimensions of *correctness* - *accuracy, completeness, correctness, consistency* and *conciseness*; all of the dimensions are evaluated in their experiment. In [26], the *correctness* of a formalization is also evaluated in the experiment, but focuses on *"whether the code does what it should and nothing more"*. The term "code" refers to logical the representation. In other cited projects, the *correctness* of a legal formalization is equal to the decision on the interpretation of a legal text as provided by a legal expert (sometimes in cooperation with a logician or a computer scientist).

Nevertheless, a single and correct logical representation of a certain legal text is difficult to find even for legal experts. Furthermore, the examples provided in the studies are usually simplified. More complicated legal texts with more complicated (or controversial) interpretations are not usually used as suitable examples. Therefore, the opinion on a possibility of finding a single *correct* interpretation of a legal text is abandoned, and the idea of different possible interpretations dependent on other legal (additional legal sources) and extralegal (time, social context etc.) circumstances is becoming the leading one [10]. This issue is further discussed in Subsection 4.4.

It should be noted that it is still necessary to define a single *correct* interpretation for specific circumstances at a particular time, so that a legal formalization can be further used for automatic legal reasoning. The authors, therefore, come up with a different approach to this parameter and use well-established principles from legal science related to legal text interpretation, as described thereof. There is a broad consensus that the legal language uses vague and ambiguous terms. On the other hand, as the law has to be respected by the general public, most cases and situations must have only one possible interpretation, which is straightforward and agreed upon by the majority of the addressees of the law. Otherwise, such a legal system would not provide legal certainty and would not be coherent and socially acceptable.

This principle is well reflected in the theory of soft cases and hard cases [19, 16]. Based on this theory, the vast majority of cases (legal conflicts) are soft or easy - they can result from the text itself or from the straightforward interpretation of the legal text. Only a small part of legal conflicts require more advanced interpretation methods and the results of the interpretations can be controversial, with several possible reasonable outcomes. The authors apply this methodology analogically to the interpretation of legal text for legal formalization. Given this theory, the vast majority of legal rules should be formalized in a non-controversial way and it should be possible to find a broad agreement on a single interpretation. However, there will always be a small part of the legal

rules, which will be problematic for formalization because of multiple possible reasonable interpretations.

A suitable evaluation method of this definition of *correctness* is then a rating of *correctness* in a broad evaluation experiment by legal experts. The closest to our suggested methodology is the one in [8]. A simple question on the *correctness* of legal formalization is suitable. However, there are some issues related to the evaluation of this question.

First, it is important to take into account the size of the evaluation group and ideally, the objectivity and expertise of the evaluators. As it was shown in Section 3, legal formalization is usually evaluated in small-scale experiments and furthermore, evaluators are very often the authors or colleagues of the authors. For the results to be significant, a larger evaluation group with independent members ideally with various expert backgrounds is necessary.

Second, it is necessary to take into account the form in which the formalization of a legal text is presented for the evaluation. Legal experts are not usually familiar with legal formulae, or with representations of legal text in code form. It is therefore necessary to provide a suitable tool for the translation of logical formulae back into a language suitable for evaluation. This issue is strongly related to the *transparency* of legal formalization and is described in the following subsection.

4.2 Transparency

Compared to the *correctness* of legal formalization, there is much less research on the *transparency* of legal formalizations (for example [23]). The *transparent* manner of the translation from legislation to logical formulae is necessary for the assessment of all of the other parameters. Mapping the logical relationships among legal terms in the original legal text and encoding them in logical formulae usually require at least two experts - a logician (or a computer scientist) and a legal expert. A very common process is the following: a legal expert provides the interpretation of a legal text, and a logician (or a computer scientist) translates this expert's interpretation into logical formulae. In such a case it is very difficult to evaluate the correctness of such formalization - because none of the experts understands perfectly both sides of the process. Recently, there are approaches of how to overcome this gap (for example in [22, 23, 30]) by using tools to provide a comprehensible one-to-one mapping of original legal text and logical formulae, making legal formalization *transparent*.

In another work [8], basic graphical methods (indentation) are used for presenting the logical formalization of the original legal text. In [26], the authors had to provide exhaustive explanation of the code as an output and its relationship to the original legal text first. However, these methods were tailored specifically for one formalized example, which was described in cited articles. Thus, it is not possible to draw general conclusions.

Suitable user interface interactively connecting the original legal text and its logical representation is definitely a step toward to a better transparency of this process. An effort in this direction is described in [25], where it is suggested to

evaluate *transparency* of a legal formalization in a similar way to the evaluation of *correctness*, e.g. asking legal experts about their understanding of the mapping between the original text and logical formulae.

4.3 Comprehensibility

The *comprehensibility* of a legal formalization is closely related to its *correctness* and its *transparency*. Although these three terms are separated, their evaluation will often overlap in practice. The *comprehensibility* of a legal formalization lies in a general understanding of the method and its result, i.e logical formulae. Where the *transparency* parameters should evaluate the relationship between the original text and its logical representation, the *comprehensibility* parameter should evaluate the complexity of the logical representation as an output. The *comprehensibility* of such an output is necessary for the evaluation of a logical formalization as well as for the broader use of the evaluated methodology. Simply put - logical formalization which is difficult to read, analyze or understand is not very suitable to be used in practice by lawyers or laymen. In this regard, this parameter is closely related to the friendliness of a user interface and the presentation of the formalization. The authors believe that a more comprehensible output of the legal formalization is a crucial step towards a wider use of the methods and large-scale evaluations and therefore, towards more significant results.

However, reviewed works rarely contain any consideration of a methodological approach (and not just *ad hoc*) to the comprehensibility of their outputs which are evaluated. In [26], the authors present logical representations to the evaluators of the code with an exhaustive text explanation of its meaning. On the other hand, the comprehensibility is one of the parameters they evaluate (in questions *"Can you read the code without getting a headache? Can you understand the code?"*) with quite satisfactory results. In [25], the authors provide an example of a tool with a user interface for legal formalization. Despite the fact that they do not have evaluation results for the use of the tool by lawyers, they provide freely available access to the tool and this tool contains a user interface suitable for the use of lawyers and laymen.

4.4 Multiple interpretations support

As it was described above, the *support of multiple interpretations* for a single legal text is necessary for several reasons. There is an extensive literature body related to the ambiguity and vagueness of legal text [15, 37] and very often the legal discourse itself does not agree on a single correct interpretation. Additionally, there are well-described legal and extralegal circumstances causing the ever-changing characteristics of the law.[19] It is very common, that generally accepted interpretation of a certain legal rule changes in the context of related higher court decisions even in continental legal systems. Furthermore, there are social changes and novelizations of legislation which change the interpretation

every now and then. Therefore, systems which are rigidly dependent on one interpretation of a legal text, which is moreover highly laborious, will always be limited for use and very probably highly maintenance intensive. To deal with this issue, logical methods supporting defeasibility were adopted in [21, 25, 8].

This situation favours systems and methods that are *dynamic*. Which means that the formalization can be easily changed or it can *support several interpretations* of single legal text at once. The authors suggest evaluating this as a further parameter of legal formalization methods. Currently, none of the reviewed works contains the evaluation of this parameter and the results of legal formalization are presented to the evaluators as they are, i.e as a single interpretation. However, with an appropriate and friendly user interface, it is advisable to give the lawyers the possibility to provide their interpretations of a certain legal text. The results from experimental legal formalization performed by lawyers themselves with the support of a suitable tool or a system may provide highly significant results in the evaluation. On the other hand, it raises the need for a comprehensible and friendly user interface for legal formalization.

5 The legal experts evaluation methodology proposal

As was suggested in the previous section, we consider the expert group's evaluation of the four presented parameters as a way to go when answering the research question posed in Section 1. As it was mentioned in the previous section, legal experts' opinion on the four parameters of certain formalization is necessary for its use in practice as there is no other authority that can decide on the interpretation of legal rules and its correct formalization into logical formulae. Furthermore, for this decision to be objective and significant, the evaluation experiment should meet the following conditions.

First, as it was mentioned in Subsection 4.1, the group of legal experts should be consisting of experts (ideally with different backgrounds depending on the goal of such experiment) independent of the research author team for the evaluation to be objective. Regarding the size of the group, the standard rule applies: the larger the group, the more significant the results.

Second, the evaluation experiment should target all of the suggested parameters as all of them are necessary for a meaningful legal formalization and are closely related to each other. Ideally, the questions asked to the evaluators should be as expressive as possible and as neutral as possible. For example, the question *"Can you read the code without getting a headache?"* from [26] would be probably assessed as misleading, which again reduces the significance of results. The rating scale should usually contain an even number of options, as it avoids selecting middle rating options [35].

Third, the group of legal experts should be divided in such a manner that inter-rater agreement is possible to measure as well. Significant differences in ratings usually lead to less significant results of the evaluation.

Fourth, recent approaches to understanding artificial intelligence and its impact on the provision of legal services (or, more generally, on the provision of any

services that were previously the exclusive domain of humans) have emphasized the interactive process and collaboration with the artificial intelligence or an AI-based tool rather than the service providing the final outcome or decision. Such an approach is more flexible and combines technology as a means of obtaining information and humans as the decision-making entity. Such an approach can also be applied to evaluation and the evaluation experiment can be seen not as a one-off evaluation but rather as a collaboration on a suitable solution, as in e.g. [1] or [36]. The disadvantages of such an approach are, of course, the greater time and technology requirements.

The proposal presented here for a methodological evaluation of legal formalization methods is definitely not a ready-to-use system for evaluation, although its aim is to suggest starting points that must be taken into account when designing a specific evaluation experiment. Subsequently, its goal is to advocate for more frequent evaluations of legal formalization methods and to provide guidance for more significant and comparable research results.

6 Conclusions

In this article, the authors assessed the question of what is a "good quality" legal formalization method. First, an overview of recent state-of-the-art research efforts in legal formalization was presented. Second, a description of how the cited works are evaluated is given and the authors discuss how the different evaluation methods can be employed in a general context. Based on this overview, necessary properties of a "good" legal formalization are identified - *correctness, transparency, comprehensibility* and *multiple interpretation support*. Lastly, the authors argue that the most suitable methodological approach to the legal formalization evaluation is a human-centered (ideally legal experts group) experiment. The suggested experiment should further focus on necessary properties of legal formalization to be meaningful and objective. In this regard, these four properties should serve successfully as parameters for objective and comparable results of future evaluations.

References

1. Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., van Hoof, H., van Riemsdijk, B., van Wynsberghe, A., Verbrugge, R., Verheij, B., Vossen, P., Welling, M.: A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer **53**(8), 18–28 (Aug 2020). https://doi.org/10.1109/MC.2020.2996587
2. Aleven, V., Ashley, K.D.: Evaluating a learning environment for case-based argumentation skills. In: Proceedings of the sixth international conference on Artificial intelligence and law - ICAIL '97. p. 170–179. ACM

Press, Melbourne, Australia (1997). https://doi.org/10.1145/261618.261650, http://portal.acm.org/citation.cfm?doid=261618.261650
3. Aleven, V.A.: Teaching case-based argumentation through a model and examples. Citeseer (1997)
4. Araszkiewicz, M., Zurek, T.: Comprehensive framework embracing the complexity of statutory interpretation. Legal Knowledge and Information Systems p. 145–148 (2015). https://doi.org/10.3233/978-1-61499-609-5-145
5. Ashley, K.D.: Modelling Legal Argument: Reasoning with Cases and Hypotheticals. Ph.D. thesis, University of Massachusetts, USA (1988)
6. Atkinson, K., Collenette, J., Bench-Capon, T., Dzehtsiarou, K.: Practical tools from formal models: the echr as a case study. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. p. 170–174. ACM, São Paulo Brazil (Jun 2021). https://doi.org/10.1145/3462757.3466095, https://dl.acm.org/doi/10.1145/3462757.3466095
7. Bartolini, C., Giurgiu, A., Lenzini, G., Robaldo, L.: Towards legal compliance by correlating standards and laws with a semi-automated methodology. In: Bosse, T., Bredeweg, B. (eds.) BNAIC 2016: Artificial Intelligence. p. 47–62. Communications in Computer and Information Science, Springer International Publishing, Cham (2017)
8. Bartolini, C., Lenzini, G., Santos, C.: An agile approach to validate a formal representation of the gdpr. In: Kojima, K., Sakamoto, M., Mineshima, K., Satoh, K. (eds.) New Frontiers in Artificial Intelligence. p. 160–176. Lecture Notes in Computer Science, Springer International Publishing, Cham (2019)
9. Bench-Capon, T.J.M., Coenen, F.P.: Isomorphism and legal knowledge based systems. Artificial Intelligence and Law **1**(1), 65–86 (Mar 1992). https://doi.org/10.1007/BF00118479
10. Branting, L.K.: Data-centric and logic-based models for automated legal problem solving. Artificial Intelligence and Law **25**(1), 5–27 (Mar 2017). https://doi.org/10.1007/s10506-017-9193-x
11. Chelimsky, E.: The coming transformations in evaluation. Evaluation for the 21st century: A handbook pp. 1–26 (1997)
12. Cohen, P.R., Howe, A.E.: How evaluation guides ai research: The message still counts more than the medium. AI magazine **9**(4), 35–35 (1988)
13. Conrad, J.G., Zeleznikow, J.: The significance of evaluation in ai and law: a case study re-examining icail proceedings. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law. p. 186–191. ICAIL '13, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2514601.2514624, https://doi.org/10.1145/2514601.2514624
14. Conrad, J.G., Zeleznikow, J.: The role of evaluation in ai and law: an examination of its different forms in the ai and law journal. In: Proceedings of the 15th International Conference on Artificial Intelligence and Law. p. 181–186. ICAIL '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2746090.2746116, https://doi.org/10.1145/2746090.2746116
15. Dworkin, R.: No right answer. NYUL Rev. **53**, 1 (1978)
16. Fischman, J.B.: How many cases are easy? Journal of Legal Analysis **13**(1), 595–656 (2021)
17. Hall, M.J.J., Hall, R., Zeleznikow, J.: A process for evaluating legal knowledge-based systems based upon the context criteria contingency-guidelines frame-

work. In: Proceedings of the 9th international conference on Artificial intelligence and law. p. 274–283. ICAIL '03, Association for Computing Machinery, New York, NY, USA (2003). https://doi.org/10.1145/1047788.1047843, https://doi.org/10.1145/1047788.1047843
18. Hall, M.J.J., Zeleznikow, J.: Acknowledging insufficiency in the evaluation of legal knowledge-based systems: Strategies towards a broadbased evaluation model. In: Proceedings of the 8th international conference on Artificial intelligence and law. pp. 147–156 (2001)
19. Kühn, Z.: Aplikace práva ve složitých případech: k úloze právních principů v judikatuře. Karolinum (2002)
20. Lawsky, S.: Form as formalization symposium on artificial intelligence and the future of tax law: Writing laws that robots can read. Ohio State Technology Law Journal **16**(1), 114–156 (2020)
21. Lawsky, S.B.: A logic for statutes. Florida Tax Review **21**(1), 60–80 (2017)
22. Libal, T.: A meta-level annotation language for legal texts. In: International Conference on Logic and Argumentation. pp. 131–150. Springer (2020)
23. Libal, T., Novotná, T.: Towards transparent legal formalization. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 296–313. Springer (2021)
24. Libal, T., Novotná, T.: Towards automating inconsistency checking of legal texts. Jusletter IT (27-Mai-2020) (2020)
25. Libal, T., Steen, A.: Towards an executable methodology for the formalization of legal texts. In: Dastani, M., Dong, H., van der Torre, L. (eds.) Logic and Argumentation. p. 151–165. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020)
26. Merigoux, D., Chataing, N., Protzenko, J.: Catala: A programming language for the law. Proceedings of the ACM on Programming Languages **5**(ICFP), 1–29 (Aug 2021). https://doi.org/10.1145/3473582, arXiv: 2103.03198
27. de Montety, C., Antignac, T., Slim, C.: GDPR Modelling for Log-Based Compliance Checking, IFIP Advances in Information and Communication Technology, vol. 563, p. 1–18. Springer International Publishing, Cham (2019)
28. Nakamura, M., Nobuoka, S., Shimazu, A.: Towards translation of legal sentences into logical forms. In: Satoh, K., Inokuchi, A., Nagao, K., Kawamura, T. (eds.) New Frontiers in Artificial Intelligence. p. 349–362. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2008)
29. Palmirani, M., Bincoletto, G., Leone, V., Sapienza, S., Sovrano, F.: Hybrid refining approach of pronto ontology. In: Kő, A., Francesconi, E., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) Electronic Government and the Information Systems Perspective. p. 3–17. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020)
30. Palmirani, M., Governatori, G.: Modelling legal knowledge for gdpr compliance checking. Legal Knowledge and Information Systems p. 101–110 (2018). https://doi.org/10.3233/978-1-61499-935-5-101
31. Pertierra, M.A., Lawsky, S., Hemberg, E., O'Reilly, U.M.: Towards formalizing statute law as default logic through automatic semantic parsing. In: ASAIL@ ICAIL (2017)
32. Prakken, H., Wyner, A., Bench-Capon, T., Atkinson, K.: A formalization of argumentation schemes for legal case-based reasoning in aspic+. Journal of Logic and Computation **25**(5), 1141–1166 (Oct 2015). https://doi.org/10.1093/logcom/ext010

33. Robaldo, L., Bartolini, C., Palmirani, M., Rossi, A., Martoni, M., Lenzini, G.: Formalizing gdpr provisions in reified i/o logic: The dapreco knowledge base. Journal of Logic, Language and Information **29**(4), 401–449 (Dec 2020). https://doi.org/10.1007/s10849-019-09309-z
34. Rotolo, A., Governatori, G., Sartor, G.: Deontic defeasible reasoning in legal interpretation: two options for modelling interpretive arguments. In: Proceedings of the 15th International Conference on Artificial Intelligence and Law. p. 99–108. ICAIL '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2746090.2746100, https://doi.org/10.1145/2746090.2746100
35. Simms, L.J., Zelazny, K., Williams, T.F., Bernstein, L.: Does the number of response options matter? psychometric perspectives using personality questionnaire data. Psychological Assessment **31**(4), 557–566 (Apr 2019). https://doi.org/10.1037/pas0000648
36. Steging, C., Renooij, S., Verheij, B.: Rationale discovery and explainable ai. Legal Knowledge and Information Systems p. 225–234 (2021). https://doi.org/10.3233/FAIA210341
37. Tushnet, M.V.: Critical legal theory. The Blackwell Guide to the Philosophy of Law and Legal Theory pp. 80–89 (2005)
38. Verheij, B.: Formalizing arguments, rules and cases. In: Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law. p. 199–208. ICAIL '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3086512.3086533, https://doi.org/10.1145/3086512.3086533