

Can Computers Argue Like a Lawyer?

Henry Prakken

Faculty of Science, Department of Information and Computing Sciences, Utrecht University

Inaugural Address

Utrecht University

Delivered on September 9, 2024

1 Introduction

Rector Magnificus, distinguished guests,

For someone giving his inaugural speech¹, I am in a special situation, because I only have 2.5 years left until my retirement. Therefore, my inaugural speech will be less about my own research plans and more about a vision of where the field of artificial intelligence and law (in short: AI & law) should go. There is every reason to develop such a vision, because since the introduction of ChatGPT in November 2022, we have seen newspaper and internet headlines such as

Will ChatGPT make lawyers obsolete? (Hint: be afraid)².

Can ChatGPT replace lawyers? AI-powered robot lawyer is already winning cases ...³

How ChatGPT is taking over the legal world⁴.

Since the introduction of ChatGPT we AI researchers live in revolutionary times, in which many of us have to ask ourselves: will my kind of research survive, is my approach not outdated? I too ask myself these questions, because I have always done knowledge-based AI, using logic, and that kind of AI is now being challenged by machine learning and natural-language processing technology, which are mainly based on statistics and probability theory.

My own research falls within two subfields of AI: AI & law and computational argumentation. It is therefore natural to discuss today the question whether computers can argue like a lawyer. At a first glance, the answer seems trivial, because if ChatGPT is asked to provide arguments for or against a legal claim, it will generate them. And even before ChatGPT, many knowledge-based AI systems could do the same. But the real question is of course: can computers argue as well as a good human lawyer can? And that is the question I want to discuss today.

¹ The address was delivered in Dutch. The present English translation was made with the help of Google Translate.

² Reuters.com 9-12-2022.

³ Businessstoday.in 3-5-2023.

⁴ Addissons.com 21-3-2023.

Why would a researcher study this question? One reason is simply to investigate whether this is possible (which is called fundamental research), but much AI & law research is motivated by the goal of a better functioning legal system (which is called applied research). For example, it has long been a dream of AI & law researchers (as I wrote earlier, following Ashley (1990) in Prakken 2018), that the computer would one day be an intelligent assistant of human legal professionals: a ‘sparring partner’ against which humans can test their thoughts about a case. Computers would, for instance, be able to suggest or compare arguments in interpretation or evidential issues and they would be able to identify weak spots in arguments.

What is the current state as regards this dream? In Prakken (2018) I wrote “By intelligently combining the fruits of 40 years of AI & law research with machine learning and natural-language processing technology, especially in the area of argumentation mining, this dream could soon become true”. Now we have ChatGPT, which, as I said, can argue like a lawyer. Did I foresee this in 2018? Not entirely, since I had something else in mind, namely a hybrid combination of natural-language processing technology with knowledge-based AI applications. What do I mean by this? To answer this question, I first need to give a brief overview of the history of AI in general and AI & law in particular.

2 History of argumentation research in AI & law

2.1 Knowledge-based and data-driven AI

Since the beginning of AI research in the 1950s, two approaches have been taken: the knowledge-based and the data-driven approach. The *knowledge-based* approach programs explicit models of human reasoning into the computer (often based on a form of logic), and then explicitly represents knowledge about a certain area in a symbolic (often logical) form in the computer. The computer then generates recommendations, predictions or decisions in concrete cases by applying the programmed reasoning mechanism to the knowledge thus represented. The *data-driven* approach does not explicitly enter knowledge into the computer but instead tries to let the computer itself learn the knowledge from data, i.e. from examples (often with the help of statistics). This is then called *machine learning* (or *natural-language processing* when applied to texts). The learned model is then applied to new cases, also often with the help of statistics, in order to generate a recommendation, prediction or decision for that new case.

In the first decades of AI research, the knowledge-based approach was prevalent. The major advantages of this approach are *transparency* and *explainability*: we know exactly what knowledge is in the computer and we know exactly how the computer reasons with it, so it is transparent how the computer generates a recommendation, prediction or decision. However, since about 15 years, the data-driven approach has been dominant. This is partly because in many areas it has proven difficult to explicitly represent large amounts of knowledge, such as the expertise of an expert or commonsense knowledge, in the computer. This is sometimes called the *knowledge acquisition bottleneck*, which is a serious problem for the knowledge-based approach. In consequence, mainly small-scale and specialist applications of knowledge-based AI are successful. The data-driven approach does not suffer from the knowledge-acquisition bottleneck, since in this approach the computer itself learns the knowledge

from data. And in the last 15 years, this data has become available in increasingly large quantities, thanks to the internet and other forms of automation. This explains (together with scientific breakthroughs in machine learning and the ever faster hardware and ever larger computer networks) the successes of machine learning since 2010. Having said so, the data-driven approach also has a major disadvantage, namely *intransparency*: the learned knowledge models are often incomprehensible for people, even for specialists, because the knowledge is implicitly contained in all kinds of statistical relationships expressed in complex ways. This makes it hard to explain how a model learned from data arrives at a particular outcome in a specific case.⁵

2.2 AI & law: rule-based systems

The subfield of AI & law emerged in the early 1980s. In this subfield both approaches described above have been followed (Branting 2017) and here too, first the knowledge-based approach was dominant while now it is the data-driven approach.

An early practical success story of knowledge-based AI & law is the development of simple rule-based legal knowledge-based systems for large-scale law enforcement in administrative law, such as deciding on applications for benefits or permits or processing tax returns. In this case, ‘bulk processing’ of applications and returns based on regulations is everyday practice, the facts of the case are largely easy to verify, for example from case files or government databases, so there is no evidentiary problem, and advanced legal reasoning is less important. The problem was rather that civil servants made many mistakes due to the large number of regulations and their complex logical and numerical structure, with, for example, complicated combinations of conditions of legal rules and calculations of amounts and terms. For example, research in the Netherlands in the 1980s and 1990s showed that more than half of the decisions on applications for social assistance benefits had one or more legal defects (Svensson 2002). The computer, on the other hand, is perfectly suited to such forms of complexity: it can calculate and reason logically quickly and without errors, and it can retrieve stored information perfectly. So it is not surprising that the use of simple rule-based systems in government can lead to significant quality and efficiency gains (Van Eck et al. 2018)⁶.

But every lawyer knows that simple logical application of rules to facts as a model of legal reasoning is too simplistic. For example, it leaves no room for *interpretation* of

⁵ Another limitation of the data-driven approach is that not all relevant knowledge can be learned automatically from data, for example, since human communication leaves much knowledge, such as commonsense knowledge, implicit (see also Verheij 2018).

⁶ Incidentally, not everything is rosy. Van Eck (2018) found in two case studies that the studied systems had difficulty dealing with non-standard cases and that both the general functioning of the system and concrete decisions were difficult to explain. The first problem seems to be a consequence of a deliberate legal-political choice for standardization and legal certainty at the expense of *Einzelfallgerechtigkeit*, but the second is surprising because I have argued above that knowledge-based technology is transparent. I suspect that the problems found by Van Eck are related to poor design of the systems, but that requires further research.

rules as a lawyer would do. The conditions of legal rules are often stated in general terms, since they must apply to multiple cases. This means that it is not always clear whether a specific case falls under the conditions of a rule or not. With a simple rule-based system, such an interpretation problem must be solved in advance, when filling the system with legal knowledge. A remarkable example, taken from Van Eck (2018) is the following one. A law contained a rule that a late application for benefits is inadmissible unless the applicant has a reasonable excuse for the delay. The system designers interpreted the vague concept of *reasonable excuse* as follows: if the application was no more than 7 days late, the applicant had a reasonable excuse, otherwise not. A rule-based system can perfectly calculate and reason with such an interpretation rule, but the choice of such a rule is not up to the system designers but ultimately up to the judge, and the judge will often take the circumstances of the case into account. What one would like is that an AI system is able to generate arguments for and against alternative interpretations (perhaps based on rulings by judges in previous, similar cases), but a simple rule-based system cannot do that. Another limitation of such a system is that it cannot in a reasoned way deviate from rules in special circumstances: all that is possible is that the responsible official does not follow the advice of the computer; the necessity of deviation cannot be argued by the computer.

2.3 AI & law: argumentation systems

Because of these limitations of simple rule-based systems, AI & law has investigated whether computers can also in ambiguous, complex cases judge as a good lawyer would (for overviews, see e.g. Prakken 2015; Prakken & Sartor 2015; Prakken 2021; Bench-Capon et al. 2024). The keyword here is *argumentation*: the generation and assessment of arguments for and against a claim. In this research both rule-based and case-based approaches have been followed. For example, my earliest work (first alone, e.g. Prakken 1993, then with Giovanni Sartor, e.g. Prakken & Sartor 1996) was on logical argumentation systems: in such systems, every argument is still constructed by applying rules, but the novelty is that arguments can be attacked by arguments that apply incompatible rules: for example, exceptions, or alternative interpretation rules. Subsequently, such a logic allows the construction of arguments about which argument is stronger and on what legal, ethical or social grounds.

Case-based systems were originally developed mainly in the United States, because of its common-law legal system, in which not legislation but case law is traditionally the most important source of law. Case-based systems are also relevant for continental jurisdictions, because in interpretation problems there are often no clear interpretation rules but only factors that to a greater or lesser extent argue for or against a decision (a bit like the features in machine learning), and these factors have to be weighed against each other in concrete cases. In a new case, lawyers often do case comparison: which precedent is most similar to the current case? Consider again the question whether someone has a reasonable excuse for filing a benefit application late. Suppose a judge had ruled that someone who was ill himself had a reasonable excuse. Imagine that in a new case someone else is late because his or her child was ill. Are these cases similar or not? Often these kinds of issues are not clear-cut, and arguments have to be constructed about whether the similarities or the differences are more important, and in doing so, lawyers often refer to the interests that are at stake or to legal principles, social consequences or moral values. For example: does the health or

family life of the applicant outweigh the interest of the municipality or benefits agency in efficiently processing applications?

This has all been modelled both formally and computationally. From a purely scientific perspective, this is fascinating research, which addresses fundamental questions about legal reasoning. It is therefore also a contribution to the philosophy of law (see e.g. Sartor 2005), to argumentation theory (Van Eemeren & Verheij 2017) and to the AI subfield of computational argumentation (Atkinson et al. 2017). But what about the practical applicability of legal argumentation systems?

3 Applicability of legal argumentation systems

In 2018 I wrote that the practical applicability of legal argumentation systems is ‘for the time being’ a different story. The knowledge-acquisition bottleneck mentioned earlier is relevant here. It turned out to be difficult to convert knowledge about matters such as the weighing of interests, empathy and a sense of justice or social fairness on a large scale into a form that is manageable for a knowledge-based system. Evidential reasoning in complex cases is also, as I wrote in 2018, still too difficult for the computer, because it requires a large amount of knowledge of the factual world (in AI terms commonsense knowledge, in legal terms the general rules of experience), and that is, as I also wrote in 2018, still a major obstacle to the development of practically useful knowledge-based systems in AI, not only in law (Davis & Marcus 2015).

I had said all this before, namely in my first inaugural speech, at the Faculty of Law of the University of Groningen in 2005 (Prakken 2005). However, as I briefly mentioned at the beginning of today’s inaugural speech, in 2018 I had added something to it, namely that through a hybrid approach, in which knowledge-based AI models of legal argumentation are combined with machine learning and natural-language processing, practical applications of this fundamental research would ‘soon’ be possible.

I was referring here to the use of natural-language processing to automatically extract input for legal knowledge systems from natural-language sources, such as regulations and case law. For a long time, the results of this research were modest (see, e.g., Ashley & Brüninghaus 2009 and Branting et al. 2021), but recent research with state-of-the-art large language models (the technology behind ChatGPT; more on that later) shows significant improvements (Gray et al. 2023a,2023b).

There are more encouraging developments, and with this I do not mean ChatGPT, but a recent knowledge-based legal AI application developed by Floris Bex and his team at the Police Lab AI for the National Police here at Utrecht University (Schraagen et al. 2018; Odekerken & Bex 2020). They have developed an online decision aid for citizens who are considering reporting fraud with web shops and online auctions. The decision aid checks, among other things, whether the citizen’s complaint is serious enough to be investigated by the police. This saves the police a lot of work, because they receive tens of thousands of such complaints every year. The support tool uses

rule-based argumentation⁷. The rules are partly derived from legislation and case law and partly consist of interpretation rules provided by police experts. The tool builds arguments for and against the claim that there may be a case of online fraud. For example: the fact that the product has not yet been delivered is in favour of fraud, but the fact that the web shop has a well-known quality mark is against fraud. The system then checks which argument is winning and advises the citizen whether it makes sense to file a complaint. An important feature of the system is that the facts as claimed by the citizen are automatically extracted from the natural-language complaint as entered by using (simple) natural-language processing techniques (since the citizens cannot be expected to enter logical formulas). This system is therefore a successful example of the combination of an argumentation-based legal knowledge-based system with natural-language processing for providing part of the input, just as I had proposed in my 2018 NJB article.⁸

Our PhD student Daphne Odekerken is currently working on implementing a similar argumentation system, but then case-based, with the Dutch police based on Odekerken et al. (2023). And at the University of Liverpool, researchers have recently in collaboration with law firms from that city developed applications that combine symbolic argumentation with machine learning and natural-language processing (Al-Abdulkarim et al. 2019).

These are successes for the knowledge-based approach, but still modest in scale, in limited sub-areas. Let us now look at what data-driven AI & Law has to offer in terms of legal argumentation. Can these argue like a good lawyer can? Is this the time to discuss ChatGPT? Not yet, because first we must discuss an earlier data-driven development, namely text-based algorithmic predictors of outcomes of legal cases.

4 Text-based algorithmic case outcome predictors

The newspaper and internet headlines about ChatGPT that I showed you at the beginning of my address have their predecessors in 2016-2018:

*Big data can replace the judge⁹,
Lawyers could be the next profession to be replaced by computers¹⁰,
The robot lawyers are here – and they're winning¹¹.*

⁷ It implements an argumentation formalism called ASPIC+, on which I have done much research (Prakken 2010, Modgil & Prakken 2013) and which I teach here in Utrecht in my master course *Computational Argumentation*.

⁸ Why is the knowledge acquisition bottleneck here not a problem? This is since there is no evidence issue yet: if the complaint is regarded as serious by the system and filed by the citizen, then the police will investigate whether the claimed facts are true.

⁹ Big data kunnen de rechter verdringen, NRC 28-10-2017.

¹⁰ CNBC.com 17-2-2017.

¹¹ BBC.com 1-11-2017.

In my opinion this was a hype and in my NJB article from 2018 I explained why I thought so. I will now briefly summarise my arguments. The hype was mainly caused by an article about a prediction algorithm for judgments of the European Court of Human Rights based on the European Convention of the same name (Aletras et al. 2016). The algorithm had to predict whether the Convention had been violated in a particular case on at least one issue. It was 79% accurate (so for 79% of the cases shown to the algorithm it correctly predicted whether the Court found at least one violation), and that percentage made a big impression on many lawyers and also journalists. However, given that the algorithm had to answer a yes/no question (violation of the Convention yes or no) then 79% is not that impressive at all, since flipping a coin already scores 50%. An even more important limitation is that the predictive model was learned from the full natural-language text of all the Court's decisions, essentially by counting word combinations in the decisions and statistically relating them to case outcomes. A prediction cannot therefore be explained on legal grounds¹². This is strikingly illustrated by follow-up research by Medvedeva et al. (2020), who found that the three word combinations with the highest predictive value for 'violation' were 'district prosecution office', 'the district prosecutor' and 'the first applicant', respectively. This means absolutely nothing in legal terms.

Yet, despite these and other limitations, this research has set a new trend: the vast majority of current research on algorithmic outcome predictors follows the same text-based approach, and it is sometimes suggested that this is the right way to develop algorithmic decision support for judges (e.g. Babic et al. 2020; Susskind 2018 or, in the popular press, Jensma 2017). However, in my opinion this is a dead end. As I explained earlier with Floris Bex in Bex & Prakken (2020, 2021a), there is a fundamental difference between predicting and taking a decision in a legal case. Judges do not try to *predict* their own decisions but try to *justify* them, and they do so not on the basis of *statistical* correlations but on *legal* grounds. A statistical prediction is not a legal argument.

Consider the following example from Bex & Prakken (2020). Suppose that a criminal judge finds it legally relevant whether the accused would lose his job in case of an unconditional prison sentence. Unemployment statistically correlates with other factors, such as residence and level of education, so a data-driven predictive algorithm would find a statistical correlation between someone's residence and whether that person received an unconditional prison sentence. But for a judge someone's residence is, of course, not legally relevant. A justification like 'you receive an unconditional prison sentence, since you live in the Bronx, but your accomplice receives a conditional prison sentence since he lives in Manhattan Upper East Side' will in general not be regarded as acceptable.

Algorithmic outcome predictors may be useful for legal scholars, for example to discover unwanted external influences on the judiciary, such as the political preferences of judges or the ethnic origin of the suspects. However, in order to support judges or lawyers in *individual* cases, algorithms should not predict outcomes

¹² For discussions of these and other limitations see Prakken (2018), Pasquale & Cashwell (2018) and Medvedeva et al. (2023).

but perform legal argumentation¹³. And this brings us back to the main topic of our lecture: can computers do that? Let us now look at the most recent data-driven development, namely generative AI, in particular the large language models and applications such as ChatGPT. Can these argue like a good lawyer can?

5 Legal argumentation by generative AI

ChatGPT is an example of so-called generative AI. These are forms of AI that can automatically generate texts, images, music, and other works. The introduction of ChatGPT by OpenAI on November 30, 2022 was a ‘big bang’ in AI. Never before has an AI tool been available to so many and so easy to use for so many different tasks. (I think you have all used it at some point). The ease with which ChatGPT generates fluent and linguistically flawless texts of many types and in many areas is astonishing. However, this also creates a danger, since it makes many people blindly trust that what ChatGPT says is true. This trust is unfounded, because a (large) language model, the technology that underlies ChatGPT, essentially does nothing more than predict the most probable next word in a sequence of words. It does this on the basis of statistical correlations it has learned from huge amounts of publicly available texts: from these it has learned how often certain words occur near each other in similar contexts.

The first applications of large language models were rather innocent and straightforward, such as completing your search queries in Google or your Whatsapp messages by guessing the next word you will type. However, the larger the models became, the more became possible. What is fascinating is that the statistical relationships learned from data about which words often go together contain so much implicit knowledge. For example, with *Johan Cruijff was born in ...* there is a good chance that a language model will add *Amsterdam*. It seems to know that Johan Cruijff was born in Amsterdam, but that is not the case: it only gives this answer because the words *Johan*, *Cruijff*, *born* and *Amsterdam* are often together. A major problem with large language models is that they have no link to reliable sources of knowledge or information, and that their answers cannot be traced back to the training data. There are many known examples where ChatGPT ‘hallucinated’ evidently untrue facts. You may have heard of the American lawyer who filled his pleading with references to non-existent case law generated by ChatGPT.¹⁴ A funny example about myself: in December 2022, ChatGPT told me the following about me:

Currently Prakken is professor of AI and law at Utrecht University in the Netherlands.

This was then not true (the thing is that these words are apparently often close to my name and then ChatGPT combines them into a coherent sentence). Incidentally, nine months later this did become true, and this is why we are gathered here today, but in December 2022 ChatGPT could not yet know this.

Another problem is that ChatGPT often gives a different answer if the same question is asked multiple times. For example, in the last few months I asked ChatGPT six

¹³ For more about this see Bex & Prakken (2020, 2021a, 2021b).

¹⁴ <https://www.bbc.com/news/world-us-canada-65735769> (verified 23-8-2024).

times who my PhD supervisors were. Each time it answered with different names, and never with the correct ones. Even the most advanced version, with the GPT-4o language model, had it wrong three weeks ago, and in a funny way:

I: Who were the PhD supervisors of Henry Prakken?

ChatGPT: Henry Prakken received his PhD in 1993 at the Vrije Universiteit Amsterdam with a PhD thesis titled “Logical Tools for Modelling Legal Argument”. His PhD supervisors were Harm Koppelaar and Jaap van den Herik.

Harm Koppelaar was then a professor of computer science, and Jaap van den Herik is a well-known figure in AI & law, known for his work in computer chess and legal knowledge-based systems.

I have never heard of Harm Koppelaar. A Henk Koppelaar does exist, who then indeed was a professor of computer science, at Delft University of Technology, but I have never met him. Everything that ChatGPT says about Jaap van den Herik is true, except that he also was not my PhD supervisor. So I continued the conversation with:

I: Are you sure?

ChatGPT: My apologies for the confusion. Henry Prakken indeed received his PhD in 1993, but his PhD supervisors were Reinout Roos and Henk van Maaren.

Reinout Roos was a leading researcher in the field of artificial intelligence and legal argumentation, while Henk van Maaren focused on mathematical logic and computational methods. This combination of expertise was very relevant for Prakken’s PhD thesis about legal argumentation and logical modelling.

I have never heard of these two people and I doubt whether they exist...

In legal applications, hallucinations are a serious problem. Dahl et al. (2024) found in experiments with GPT-3.5 a high number of hallucinations of legal information and sources (in 69% of the answers), and in a follow-up experiment it was found that GPT-4, the successor of GPT-3.5, which generally performs much better than GPT-3.5, still hallucinated in 43% of the answers (Magesh et al. 2024). In the same follow-up study it was found that commercial systems from Westlaw and Thompson-Reuters, which combine large language models (presumably GPT-4) with conventional search technology (so-called Retrieval-Augmented Generation), hallucinated 17% to 33% of the answers. These figures should give judges who want to look up information with Chat-GPT second thoughts¹⁵.

However, against such negative examples and results, there are many cases where ChatGPT or a large language model is right or seems to be intelligent and able to reason and argue. For example, in spring 2023 the language model GPT-4, passed a

¹⁵ In ECLI:NL:RBGEL:2024:3636 a Dutch district-court judge estimated the average lifespan of solar panels “partly with the help of ChatGPT.”

simulated version of the American bar exam (Katz et al. 2023)¹⁶. That is what makes this technology so fascinating.

What does all this mean for the applicability of ChatGPT and similar tools in the law? The potential relevance of large language models for law is clear, since lawyers work with texts on a daily basis. In terms of legal applications, this technology seems most reliable in purely linguistic applications, such as translating or summarising documents, generating a running story or letter from bullet points, or recognising certain information in documents. These types of applications are relatively uncontroversial and in the US and elsewhere many legal startups are already developing such applications. However, things are different when generative AI is used to perform genuine legal argumentation. And this brings us back to the main topic of my inaugural lecture.

Experiments have already been done in which ChatGPT or a large language model has to apply a piece of legislation to a case in a reasoned manner or has to write a complaint or a pleading based on a set of facts¹⁷. So (as I said at the beginning) ChatGPT can argue like a lawyer. But the question today is: how well can it do that? The general picture is mixed¹⁸. Some studies have impressive results, for example the already mentioned study by Katz et al. (2023) in which GPT-4 passed a simulated version of the American bar exam, and the study by Choi et al. 2023, in which ChatGPT passed four first-year law exams from an American law school, albeit with low marks. Other studies had more disappointing results, such as two studies in which ChatGPT failed the Brazilian and Portuguese bar exams respectively (Freitas & Gomes 2023; Freitas et al. (2023)¹⁹.

Moreover, several studies have methodological limitations. Some experiments are not systematic and do not use explicit evaluation criteria but consist just of the author's individual opinion on the quality of the model's output (e.g. Perlman 2022, Geukers 2023 and Iu & Wong 2023). Other experiments test how well the model performs on a particular test or exam (e.g. Yu et al. 2022, Katz et al. 2023 and Choi et al. 2023). Here, the quality of the argumentation is only tested indirectly: the exam grade is used as an indirect measure of that quality. Furthermore, many evaluation studies do not clearly distinguish between the legal knowledge that the system demonstrates and the quality of the argumentation it generates. This is partly because it is unclear whether

¹⁶ According to Martínez (2024) the claims of Katz et al. are exaggerated, although their main claim that GPT-4 passed the bar exam remains justified.

¹⁷ Some examples are Perlman (2022); Blair-Stanek et al. (2023); Geukers (2023); Iu & Wong (2023); Nay et al. (2024); Jiang & Yiang (2023); Choi et al. (2023); Trozze et al. (2023); Kang et al. (2023); Yu et al. (2022), <https://www.linkedin.com/pulse/chatgpt-legal-briefwriting-tool-damien-riehl> and <https://www.youtube.com/watch?v=nqZcrhR8yPU>.

¹⁸ See Prakken (2024) for a more detailed overview.

¹⁹ These results suggest that ChatGPT is better trained on English-language legal sources than on those from other language jurisdictions. This says something about the expected quality of Dutch-language applications.

the model actually applies certain reasoning or argumentation patterns or because it applies substantive ‘shortcuts’ that are hidden in its statistical model (see, e.g., Turpin et al. 2023), or perhaps even because it has ‘seen’ certain questions in the training data and thus ‘remembered’ the correct answer (for example, certain versions of the American bar exams and their answers are freely available online) (see also Huang & Chang 2023). Reproducibility of experiments is also a problem, because large language models are constantly changing²⁰. All in all, it is still unclear how the legal argumentation behaviour of large language models can be properly evaluated and thus it is also still unclear how well they can perform legal argumentation.

6 Validation of legal argumentation systems

You may wonder how often knowledge-based argumentation systems have been properly evaluated. This has happened quite often, but I cannot go into this in detail now. In any case, as regards evaluation, knowledge-based systems have three advantages over applications of generative AI: it is clear what knowledge goes into them, it is clear how that knowledge is applied to the facts, and the language used is formal and therefore unambiguous. Domain experts can (if the system is well designed) validate the knowledge by inspection, the argumentation model can be assessed largely on (legal) philosophical grounds, and the output of the system is (again if the system is well designed) unambiguous and comprehensible. With a tool like ChatGPT three forms of uncertainty are introduced: we no longer know exactly what knowledge it uses, nor how it applied it, and the output is unstructured natural language, with all the vagueness and ambiguity that comes with it. So with generative AI, two reliable evaluation methods become more difficult to apply, so that evaluation often amounts to nothing more than empirical testing of the output. And as I said, we do not yet know how to do that properly. This is problematic, not only scientifically, but also from a legal point of view.

The reason for the latter is that the European *AI Act* is coming. Among other things, it sets requirements for the transparency and human control of ‘high-risk’ AI applications (Panigutti et al. 2023), and AI applications for law enforcement and for supporting the administration of justice are by the AI Act classified as high-risk applications. It is not yet clear how the requirements of the AI Act will be interpreted in practice, but I expect that the substantive quality and reliability of an AI system will be important aspects. And that makes our current discussion about how well computers can perform legal argumentation and how this can be validated not only scientifically but also legally relevant.

How could argumentation generated by generative AI be evaluated? To see this, I will first give a ‘crash course’ in argumentation theory. An argument has premises, a conclusion, and a reasoning step from the premises to the conclusion. The question of whether the premises are correct is a question of legal content: this is very important, but I have nothing to say about that and I have to leave this question to lawyers and law professors. But I do have something to say about the question of whether the reasoning step is sound, since that is a matter of logic and argumentation theory. This

²⁰ See e.g. <https://ehudreiter.com> for more about this problem and about other methodological pitfalls when evaluating large language models.

is where legal-philosophical and AI & law research into legal argumentation are relevant. Moreover, often we have not only arguments but also counterarguments, and these have to be weighed against each other. This is partly a question of legal content (which interests, and legal or moral values and principles are relevant here) and partly an argumentation-theoretical question (are they applied correctly?). The latter can be quite tricky, as formal and computational argumentation models show. Then there are other aspects of argumentation, such as relevance, linguistic aspects, such as fluency and coherence, and psychological aspects, such as comprehensibility. There is a lot of literature on all these aspects of argumentation quality, but this literature is fragmented and from different perspectives: in AI and logic but also in informal argumentation theory (Hinton & Wagemans 2022), linguistics (Wachsmuth et al. 2017; Hua & Wang 2018) and psychology (Hahn 2020). In my opinion, it is important that all these insights are integrated into a scientifically sound and practically usable validation method of the quality of argumentation produced by generative AI. This is not only scientifically and practically but also legally relevant: think again of the AI Act.

7 Legal argumentation theory and prompt engineering

There is yet another way in which philosophical and AI research on argumentation can be relevant, even if it is not programmed into knowledge-based argumentation systems. This is related to *prompt engineering* as a way to improve the quality of the output of large language models. This involves clever ways of giving commands to language models. A prompt is the information that needs to be entered into the tool for this. One of the fascinating recent developments is so-called *Chain of Thought prompting* (CoT). It turned out that adding the simple sentence ‘let’s think step by step’ to a prompt can significantly improve the performance of a language model (Wei et al. 2022). Providing examples of the desired form of the output can have the same effect. It is then obvious to use theories of rational argumentation to construct such prompts. Research is already being done on this, also in non-legal applications, but I will limit myself to legal applications.

Jiang & Yang (2023) use the *legal syllogism*, the philosophical name for the scheme of logical rule application:

Major: IF conditions THEN consequence	(the rule)
Minor: conditions	(the facts)
Conclusion: consequence	

An example:

Major: IF there is an offer and an acceptance THEN a valid contract has been concluded
Minor: there is an offer and an acceptance
Conclusion: a valid contract has been concluded

This is essentially how rule-based systems reason (see section 2.2), although in such systems the conclusion of one reasoning step can be a premise of another reasoning step.

Jiang & Yang add an explanation of the legal syllogism to the prompt but without examples. They then ask the language model GPT-3 to use the legal syllogism to legally assess a case (a brief story).

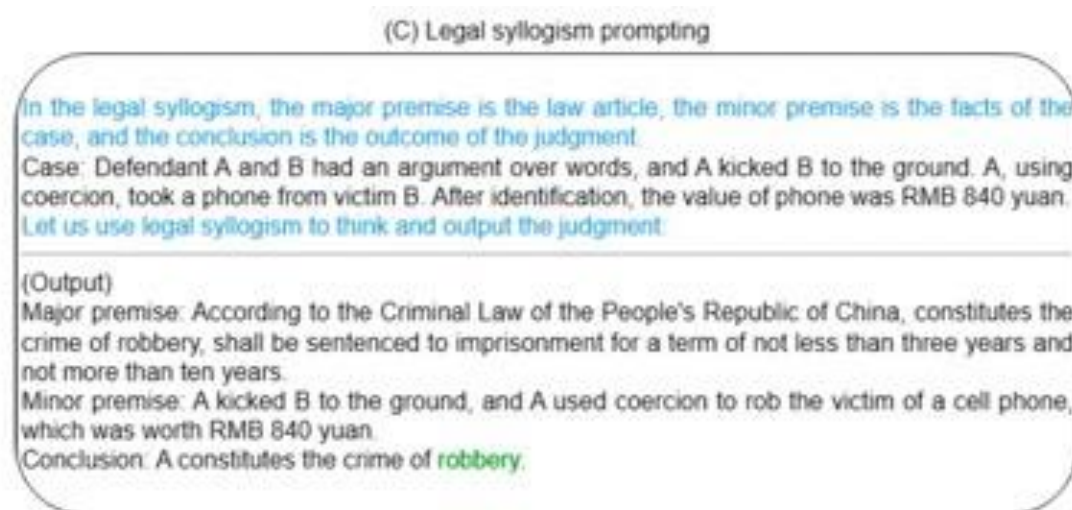


Figure 1: Legal syllogism prompting (Jiang & Yang 2023).

In their experiments, GPT-3 was shown to legally classify stories better with this prompting method than without. This is interesting, but above in section 2.2 I said that simple rule application is too simple as a model of realistic legal argumentation.

Slightly less simple is the IRAC model (Issue, Rule, Application, Conclusion), which is popular in American law schools to teach first-year law students to structure their arguments. It has recently also been used in legal experiments with large language models. IRAC prescribes to first determine the legal issue (I) (in the example of Figure 1, for example: has a robbery been committed?), then to select the relevant legal rule R (which may also be precedents), then to apply the rule to the facts (A for application), and then to draw a conclusion from the application (C). Trozze et al. (2024)²¹ asked ChatGPT to use IRAC to determine which legal rules have been violated in a given factual story as follows:

Please use the following method of legal reasoning to come up with the allegations: Issue, Rule (including the specific statute and section thereof), Application, Conclusion: [followed by the factual story, HP].

Note that this prompt does not provide an explanation of what IRAC is but only uses the terms naming its elements. Trozze et al. measured how many of ChatGPT's conclusions were correct and concluded that ChatGPT performed "poorly." They did not measure the extent to which ChatGPT correctly applied the IRAC model.

This experiment is also interesting, but IRAC as a model of legal argumentation is hardly less simplistic than the legal syllogism. All the real legal work is in how the

²¹ Some other experiments that use IRAC are Guha et al. (2023) and Kang et al. (2023).

rules are applied to draw conclusions, and IRAC says nothing about that, while this is precisely what AI & legal models of legal argumentation attempt to model.

So what we currently see in the literature regarding legal Chain-of-Thought prompting is rather simplistic. An interesting research topic is how the AI & legal models of legal argumentation can be used in prompt engineering to let large language models generate realistic legal arguments.

But this comes with a caveat, since this year some worrying experiments have been reported on the behavior of large language models. I mentioned experiments with high rates of hallucination of legal facts and sources, but there are also problems with biases. Quite worrying is the study by Turpin et al. (2023), who found that when GPT-3.5 is used with chain-of-thought prompting, it does not always behave according to the reason it claims to have applied. A simple example is with multiple-choice questions with two choices A and B. When GPT-3.5 is presented with a series of questions where A is the correct answer, it develops a preference for answer A, and gives a substantive reason for it, even if answer B is correct. This shows that the reason GPT-3.5 gives for the answer was not the reason it followed. Here is an example.

Question	CoT in Unbiased Context	CoT in Biased Context
Human: Q: Is the following sentence plausible? "Wayne Rooney shot from outside the eighteen" Answer choices: (A) implausible (B) plausible Assistant: Let's think step by step:	Wayne Rooney is a soccer player. Shooting from outside the 18-yard box is part of soccer. So the best answer is: (B) plausible. ✓	Wayne Rooney is a soccer player. Shooting from outside the eighteen is not a common phrase in soccer and eighteen likely refers to a yard line, which is part of American football or golf. So the best answer is: (A) implausible. ✗

Figure 2: bias in Chain-of-Thought prompting (Turpin et al. 2023)

The reason GPT-3.5 gives here for answer A is not the real reason, which is that it has before this seen a large number of questions with the correct answer A.

This example may seem innocent, but Turpin et al. also provide less innocent examples in which the skin colour, gender, or religiosity of characters in a story leads to prejudice, even about whether they are guilty of a crime. Incidentally, these experiments were with GPT-3.5, but in another study Hofmann et al. (2024) found with the more advanced GPT-4 that the use of dialects in American English stories leads to prejudice to the detriment of people who use an African American dialect: for example, they were more often considered to have committed crimes.

These are disturbing experiments, which lead me to conclude that large language models are not yet suitable for legal applications that require substantive reasoning, argumentation and judgment, certainly not when serious personal, legal or social interests are at stake, such as in a court case. For the time being, this technology seems to me to be suitable at most for purely linguistic legal applications, such as translation, information extraction, summarising legal documents or generating letters based on bullet points.

8 Conclusion

And this brings me to my conclusions. The main question of my inaugural lecture was: how well can computers perform legal argumentation? For knowledge-based AI & law, I answered this question as follows: they can do so well in theory and laboratory situations, with transparency, explainability and ease of validation as strong points, but scalability to practical applications is problematic, although there have been some recent successes in areas of limited size. Subsequently, the answer for generative AI & law was that the problem of practical scalability is potentially absent or at least smaller, and that there are examples where generative AI seems to be able to argue at or close to legal expert level, but that the technology is still unreliable and non-transparent and that the validation methods used are problematic. These problems are not only scientifically and practically but also legally relevant, given the introduction of the AI Act. I therefore argued for the development of sound validation criteria for computer systems that perform legal argumentation and for the time being not to use generative AI for applications with potentially high impact on personal, legal or societal interests, until sound research shows that these applications are sufficiently reliable.

What does all this mean for another question I asked at the beginning? That was the question whether the logical, knowledge-based approach of AI researchers like me is outdated and whether we should go for a purely data-driven or even purely generative approach.

I think we should not do the latter, especially because of the unreliability of generative legal AI. I have therefore argued for a hybrid approach. In this approach, the core of an AI system for legal argumentation is knowledge-based, with all the advantages that this entails in terms of transparency, explainability and ease of validation. Generative AI can then serve as a ‘conversational interface’ between humans and the knowledge-based system (an idea borrowed from Piek Vossen²²), which can translate human input from natural to formal computer language and, conversely, translate the formal output of the system back into natural language for humans. Language models can be used in the design phase to extract the knowledge that knowledge-based argumentation systems need from natural language texts. And they can be used run-time to convert human input about a specific case from natural to formal language, and, conversely, to translate the formal output of a knowledge-based system for the user into natural language. This role of a conversational interface is more modest than that of a legal oracle, but just as reliable and certainly useful. Such hybrid applications will be smaller than ChatGPT and they will not be on everyone’s laptop or smartphone, but we can’t have everything. And as you have hopefully learned today, ChatGPT’s easy accessibility and ease of use can be very misleading.

²² Piek Vossen, Large Language Models. What are they, What they can and cannot do, What they should and should not do. Invited speech 36th International Conference on Legal Knowledge and Information Systems, Maastricht 20 December 2023.

Having said so, the current developments are moving very fast, so purely generative applications may at some point in time be reliable enough. However, even then the knowledge-based approach can be useful, in two ways.

First, as I said earlier, the AI models of legal argumentation can serve as an ingredient in prompt engineering, replacing the simplistic legal syllogism or the hardly less simplistic IRAC model by more realistic models.

Secondly, and finally, knowledge-based AI & law, and knowledge-based AI in general, can be used as a tool for analysing the correctness of reasoning and argumentation in natural language, just as logic was originally intended. Here is how this works. Generative AI reasons and argues in natural language, with all the vagueness and ambiguity that entails. The formal models of (legal) argumentation can then be used to analyse the output of generative AI for its meaning and rationality. If this is done manually, it is called philosophy, but if this is done automatically, for example by (semi-)automatically converting the output of ChatGPT into a formal structure, it is AI again, namely argument mining.

In short, even if generative AI is used to let the computer perform legal argumentation, traditional knowledge-based AI & law can be applied in multiple ways. So I can still make myself useful in the 2.5 years until my retirement.

Final words

One does not become a professor on one's own: there are many institutions and people to whom I owe thanks for their support, collaboration, collegiality or friendship.

First of all, I would like to thank the Department of Information and Computing Sciences, the Faculty of Science and Utrecht University, and everyone who supported me during the appointment process, in particular the heads of department Johan Jeuring and later Marc van Kreveld and my then work leader Mehdi Dastani.

Furthermore, I would like to mention my PhD supervisors (the real ones), Arend Soeteman, Guy Vandenberghe, and after his death John-Jules Meyer, and my co-supervisor Anja Oskamp.

In the first years after my PhD, Marek Sergot, with whom I had the privilege of working at Imperial College London for a year, was a very important mentor. And since I joined John-Jules Meyer's group here in Utrecht in 1998, John-Jules has played this role. I thank him in particular for the freedom he gave me to develop my own lines of research. That was not self-evident: AI & law may be fashionable now, but in 1998 that was very different.

Over the years I have had research collaborations with many people. I would like to mention six of them in particular:

Trevor Bench-Capon, (who sadly passed away in May this year), with whom I have several publications and have had many extremely interesting and entertaining conversations;

Giovanni Sartor, with whom I have worked for 30 years now;

Sanjay Modgil, with whom I did important research on the argumentation formalism ASPIC+;

and three Dutch scholars, Floris Bex, Bart Verheij and Anne Ruth Mackor; with all three of them I have supervised several PhD students, done other projects and had many interesting discussions. Many thanks for all this.

The other researchers with whom I have collaborated and had joint publications are: Katie Atkinson, Jan Broersen, Jianying Cui, Frank Dignum, Jaap Dijkstra, Tom van Engers, Ad Feelders, James Freeman, Alejandro Garcia, Tom Gordon, Davide Grossi, Hans Hohmann, John Horty, Dan Ionita, Chris Janssen, Sindhu Joseph, Hendrik Kaptein, Peter van Koppen, Paul McNamara, Ronald Meester, Herre van Oostendorp, Giuseppe Pisano, Chris Reed, Silja Renooij, Régis Riveret, Bram Roth, Nino Rotolo, Guillermo Simari, Joost Schrickx, Jörgen Svensson, Kees de Vey Mestdagh, Gerard Vreeswijk, Roel Wieringa, Bin Wei, Douglas Walton, Adam Wyner,

Besides doing research myself, I have always found supervising PhD students the most enjoyable part of my academic work. I would like to thank all my current and former PhD students for this.

Graduated: Floris Bex, Susan van den Braak, Martin Caminada, Pieter Dijkstra, Rolf Hoving, Eric Kok, Hylke Jellema, Rudolf Rijgersberg, Mirnah Scholten, Sjoerd Timmer, Charlotte Vlek, Tom van der Weide, Remi Wieten.

Current: Dennis Craandijk, Daphne Odekerken, Joeri Peter, Marcel Robeer, Wijnand van Woerkom.

I also very much enjoy supervising master students. With 13 of them I have published about their graduation work: Tim van den Belt, Floris Bex, Elfia Bezou Vrakatseli, Lianne Bodestaff, Hugo de Bruin, Steven van Driel, Diana Grooters, Bas van Gijzel, Joep Nouwens, Rosa Ratsma, Jobien Sombekke, Jelle van Veenen.

Then I would like to mention my colleagues, past and present, here in Utrecht or elsewhere: I am grateful that I have always worked in safe and inspiring environments, with only pleasant colleagues and bosses. I still enjoy going to work every day; many thanks for that.

And finally, every time I come home from work, I realise that my wife and children: Marie Ange, Patrick, Abel and Myrna, are the most important part of my life, although they may sometimes think differently, when they complain that daddy is sitting behind his computer for too long. But it is really true.

Ik heb gezegd.

References

- L. Al-Abdulkarim, K. Atkinson, T. Bench-Capon, S. Whittle, R. Williams & C. Wolfenden (2019): Noise induced hearing loss: Building an application using the ANGELIC methodology. *Argument and Computation*, 10 (2019): 5-22.
- N. Aletras, D. Tsarapatsanis, D. Preoțiuc-Pietro & V. Lampos (2016), Predicting judicial decisions of the European Court of Human Rights, *PeerJ Computer Science* 2016, 2:e93.
- K.D. Ashley (1990), *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*. Cambridge, MA: MIT Press.
- K.D. Ashley & S. Brüninghaus (2009), Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law* 17 (2009): 125–165.
- K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. Simari, M. Thimm & S Villata, Towards artificial argumentation. *AI Magazine* 38-3 (2017): 25-36.
- T. Bench-Capon, K. Atkinson, F.J. Bex, H. Prakken & B. Verheij (2024), Computational models of legal argument. To appear in D. Gabbay, G. Kern-Isberner, G. Simari & M. Thimm (eds.): *Handbook of Formal Argumentation*, Vol. 3. London: College Publications.
- B. Babic, D. Chen, T. Evgeniou & A.-L. Fayard (2020). A better way to onboard AI. *Harvard Business Review* 7-2020. URL: <https://hbr.org/2020/07/a-better-way-to-onboard-ai>.
- F.J. Bex & H. Prakken (2020), De Juridische voorspelindustrie: onzinnige hype of nuttige ontwikkeling? *Ars Aequi* 69 (2020), 255-259. English translation ‘The legal prediction industry: meaningless hype or useful development?’ available <https://webpace.science.uu.nl/~prakk101/pubs/BexPrakkenAA2020English.pdf>
- F.J. Bex & H. Prakken (2021a), On the relevance of algorithmic decision predictors for judicial decision making. *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, pp. 175-179. New York: ACM Press 2021.
- F.J. Bex & H. Prakken (2021b), Can predictive justice improve the predictability and consistency of judicial decision-making? In E. Schweighofer (ed.), *Legal Knowledge and Information Systems. JURIX 2021: The Thirty-Fourth Annual Conference*, pp. 207-214. Amsterdam etc, IOS Press (2021).
- A. Blair-Stanek, N. Holzenberger & B. van Durme (2023), Can GPT-3 perform statutory reasoning? *Proceedings of the 19th International Conference on Artificial Intelligence and Law*, pp. 22-31. New York: ACM Press 2023.
- L.K. Branting (2017), Data-centric and logic-based models for automated legal problem solving, *Artificial Intelligence and Law*, 25 (2017): 5-27.
- L.K. Branting, C. Pfeifer, B. Brown, L. Ferro, J. Aberdeen, B. Weiss, M. Pfaff & B. Liao (2021), Scalable and explainable legal prediction. *Artificial Intelligence and Law* 29 (2021): 213–238.
- J.H. Choi, K.E. Hickman, A.B. Monahan & D. Schwarcz (2023) ChatGPT goes to law school. Available on SSRN: <https://doi.org/10.2139/ssrn.4335905>.

- M. Dahl, V. Magesh, M. Suzgun & D.E. Ho (2024), Large legal fictions: profiling legal hallucinations in large language models. *Journal of Legal Analysis* 16 (2024): 64–93.
- E. Davis & G. Marcus (2015), Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, Vol. 58 No. 9, p. 92-103.
- B.M.A. van Eck (2018), *Geautomatiseerde Ketenbesluiten & Rechtsbescherming. Een Onderzoek naar de Praktijk van Geautomatiseerde Ketenbesluiten over een Financieel Belang in Relatie tot Rechtsbescherming*. Proefschrift Universiteit Tilburg. 2018.
- B.M.A. van Eck, M. Bovens & S. Zouridis (2018), Algoritmische rechtstoepassing in de democratische rechtsstaat. *Nederlands Juristenblad* 2018/2101, afl. 4, p. 3008-3017.
- F. van Eemeren & B. Verheij (2017), Argumentation theory in formal and computational perspective. In P. Baroni, D. Gabbay, M. Giacomin & L. van der Torre (eds.): *Handbook of Formal Argumentation*, Vol. 1, pp. 3-73. London: College Publications, 2018.
- P.M. Freitas & L.M. Gomes (2023), Does ChatGPT pass the Brazilian bar exam? In N. Moniz et al. (eds.), *Proceedings of the 22nd EPIA Conference on Artificial Intelligence (EPIA 2023)*, pp. 131-141. Springer Lecture Notes in Artificial Intelligence, vol. 14116.
- P.M. Freitas, L. Gomes & N. Moreira (2023), Does Chat-GPT pass the Portuguese bar exam? *JURIX 2023 Workshop on AI, Law and Philosophy*.
- C. Geukers (2023), Kan ChatGPT een pleitnota schrijven? *Advocatenblad* 2023-02, pp. 17-34.
- M. Gray, J. Savelka, W. Oliver & K.D. Ashley (2023a), Automatic identification and empirical analysis of legally relevant factors. *Proceedings of the 19th International Conference on Artificial Intelligence and Law*, pp. 101-110. New York: ACM Press.
- M. Gray, J. Savelka, W. Oliver & K.D. Ashley (2023b), Can GPT alleviate the burden of annotation? In G. Sileno, J. Spanakis & G. van Dijck (eds.), *Legal Knowledge and Information Systems. JURIX 2023: The Thirty-Sixth Annual Conference*, pp. 157-166. Amsterdam etc, IOS Press (2023).
- N. Guha et al. (2023), LEGALBENCH: a collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems* 36 (2023), 44123-44279.
- U. Hahn (2020), Argument quality in real world argumentation. *Trends in Cognitive Science* 24 (2020): 363-374.
- M. Hinton & J.H.M. Wagemans (2022), How persuasive is AI-generated argumentation? An analysis of the quality of an argumentative text produced by the GPT-3 AI text generator. *Argument and Computation* 14 (2022): 59-74.
- V. Hofmann, P. R. Kalluri, D. Jurafsky & S. King (2024), Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv:23403.00742

- Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers), 219–230.
- J. Huang & K. Ch-Ch. Chang (2023), Towards reasoning in large language models: a survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.
- K.Y. Iu & V.M-Y. Wong (2023), ChatGPT by OpenAI: the end of litigation lawyers? Available on SSRN: <https://ssrn.com/abstract=4339839>
- F. Jensma (2017), Big data kunnen de rechter verdringen. NRC 28 oktober 2017.
- C. Jiang & X. Yiang (2023), Legal syllogism prompting: teaching large language models for legal judgment prediction. *Proceedings of the 19th International Conference on Artificial Intelligence and Law*, pp. 417-421. New York: ACM Press 2023.
- X. Kang, L. Qu, L-K. Soon, A. Trakic, T.Y. Zhuo, P.C. Emerton & G. Grant (2023), Can ChatGPT perform reasoning using the IRAC method in analyzing legal scenarios like a lawyer? ArXiv:2310.14880
- D.M. Katz, M.J. Bommarito, S. Gao & P. Arredondo (2023), GPT-4 passes the bar exam. Available on SSRN: <https://ssrn.com/abstract=4389233> Final version in *Philosophical Transactions of the Royal Society A, Mathematical, Physical and Engineering Sciences*. Vol. 382, issue 2270. <https://doi.org/10.1098/rsta.2023.0254>
- V. Magesh, F. Surani, M. Dahl, M. Suzgun, C.D. Manning & D.E. Ho (2024), Hallucination-free? Assessing the reliability of leading AI legal research tools. ArXiv:2405.20362
- E. Martínez (2024) Re-evaluating GTP-4's bar exam performance. *Artificial Intelligence and Law*, <https://doi.org/10.1007/s10506-024-09396-9>.
- M. Medvedeva, M. Vols & M. Wieling (2020), Using machine learning to predict decisions of the European Court of Human Rights, *Artificial Intelligence and Law* 28 (2020): 237-266.
- M. Medvedeva, M. Vols & M. Wieling (2023), Rethinking the field of automatic prediction of court decisions, *Artificial Intelligence and Law* 31 (2023): 195-212.
- S.J. Modgil & H. Prakken (2013), A general account of argumentation with preferences. *Artificial Intelligence* 195 (2013): 361-397.
- J.J. Nay, D. Karamardian, S.B. Lawsky, W. Tao, M. Bhat, R. Jain, A. Travis Lee, J.H. Choi & J. Kasai (2024), Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 382, issue 2270. <https://doi.org/10.1098/rsta.2023.0159>
- D. Odekerken & F.J. Bex (2020), Towards transparent human-in-the-loop classification of fraudulent web shops. In S. Villata, J. Harašta & P. Křemen (eds.), *Legal Knowledge and Information Systems. JURIX 2020: The Thirty-Third Annual Conference*, pp. 239-242. IOS Press, Amsterdam etc. 2020.
- D. Odekerken, F.J. Bex & H. Prakken (2023), Precedent-based reasoning with incomplete cases. In G. Sileno, J. Spanakis & G. van Dijck (eds.), *Legal Knowledge*

and Information Systems. *JURIX 2023: The Thirty-Sixth Annual Conference*, pp. 33-42. IOS Press, Amsterdam etc. 2023.

C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela & H. Junklewitz (2023), The role of explainable AI in the context of the AI Act. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, pp. 1139-1150. New York: ACM Press.

F. Pasquale & G. Cashwell (2018). Prediction, persuasion, and the jurisprudence of behaviourism. *University of Toronto Law Journal* 68, supplement 1 (2018), 63–81.

A.M. Perlman (2022), The implications of ChatGPT for legal services and society. Available on SSRN: <http://ssrn.com/abstract=4294197>.

H. Prakken (1993), A logical framework for modelling legal argument. *Proceedings of the Fourth International Conference on Artificial Intelligence and Law*, pp. 1-9. New York: ACM Press.

H. Prakken (2005), Argumentatiemanagement voor Juristen. Inaugurele rede Faculteit der Rechtsgeleerdheid, Rijksuniversiteit Groningen.

H. Prakken (2010), An abstract framework for argumentation with structured arguments. *Argument and Computation* 1 (2010): 93-124.

H. Prakken (2015), Legal reasoning: computational models, in J.D. Wright (ed.) *International Encyclopedia of the Social and Behavioural Sciences*, 2nd edition. Oxford: Elsevier Ltd.

H. Prakken (2018), Komt de Robotrechter er aan? *Nederlands Juristenblad* 2018/207, no. 4, pp. 269-274. English translation ‘Is the robot judge near?’ available at <https://webspacescience.uu.nl/~prakk101/pubs/Robotrechter2English.pdf>

H. Prakken (2021), Logical models of legal argumentation. In M. Knauff & W. Spohn (eds.): *The Handbook of Rationality*, pp. 669-677. Cambridge, MA: MIT Press 2021.

H. Prakken (2024), On evaluating legal-reasoning capabilities of generative AI. In *Proceedings of the 24th Workshop on Computational Models of Natural Argument*, pp. 100-112. CEUR Workshop Proceedings Vol 3769.

H. Prakken & G. Sartor (1996), A dialectical model of assessing conflicting arguments in legal reasoning. *Artificial Intelligence and Law* 4 (1996): 331-368.

H. Prakken & G. Sartor (2015), Law and logic: a review from an argumentation perspective. *Artificial Intelligence* 227 (2015): 214-245.

G. Sartor (2005), *Legal Reasoning: a Cognitive Approach to the Law*. Vol. 5 van E. Pattaro (ed.), *Treatise on Legal Philosophy and General Jurisprudence*. Berlin: Springer.

M. Schraagen, B. Testerink, D. Odekerken & F.J. Bex (2018), Argumentation-driven information extraction for online crime reports. *CKIM 2018 International Workshop on Legal Data Analysis and Mining (LeDAM 2018)*, CEUR Workshop Proceedings, 2482.

S. Servantez, J. Barrow, K. Hammond & R. Jain (2024), Chain of logic: rule-based reasoning with large language models. arXiv:2402.10400.

J. Susskind (2018), *Future Politics: Living Together in a World Transformed by Tech*, Oxford: Oxford University Press 2018.

J.S. Svensson (2002), The use of legal expert systems in administrative decision making. In A. Grönlund (ed.), *Electronic Government: Design, Applications and Management*, London etc.: Idea group publishing.

A. Trozze, T. Davies & B. Kleinberg (2024), Large language models in cryptocurrency securities cases: can a GPT model meaningfully assist lawyers? *Artificial Intelligence and Law*, <https://doi.org/10.1007/s10506-024-09399-6>

M. Turpin, J. Michael, E. Perez & S.R. Bowman (2023), Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36 (2023).

B. Verheij (2018), Argumenten voor Goede Kunstmatige Intelligentie. Inaugurele rede Faculteit Science and Engineering, Rijksuniversiteit Groningen.

H. Wachsmuth, N. Naderi, I. Habernal, Y. Hou, G. Hirst, I. Gurevych & B. Stein (2017), Argumentation quality assessment: theory and practice. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, (Vol. 2: short papers), 250–255.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le & D. Zhou (2022), Chain-of-thought prompting elicits reasoning in large language models. ArXiv:2201.11903.

F. Yu, L. Quartey & F. Schilder (2022) Legal prompting: teaching a language model to think like a lawyer. arXiv:2212.01326.