A TYPOLOGY OF LEGAL TECHNOLOGIES

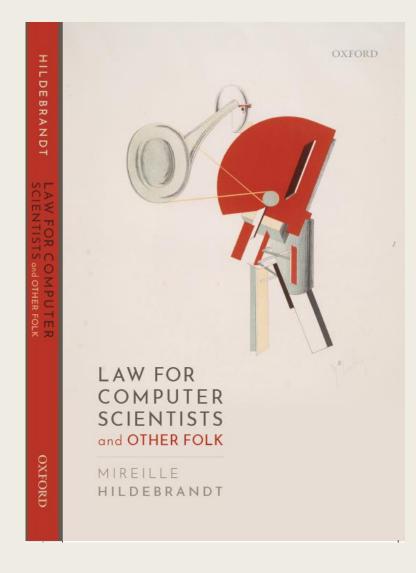
Mireille Hildebrandt, FBA
PI COHUBICOL ERC ADG project

What's next?

- Cross-disciplinary perspectives on computational 'law'
- Typology: objectives
- Typology: demonstration
- Typology: a method, a mindset beyond legal technologies

What's next?

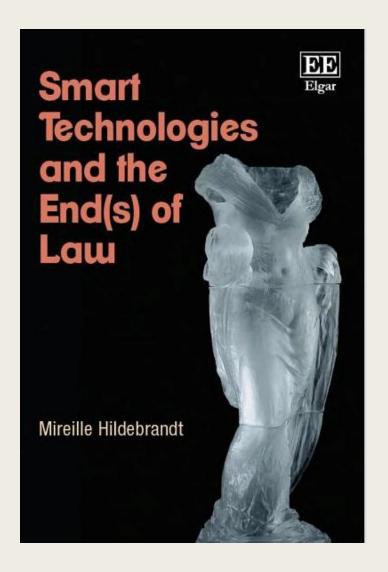
- Cross-disciplinary perspectives on computational 'law'
- Typology: objectives
- Typology: demonstration
- Typology: a method, a mindset beyond legal technologies



- My background: law, philosophy of technology
- Chair at Computer Science Department
 @Radboud University
- Research Chair at Faculty of Law & Criminology
 @Vrije Universiteit Brussel
- My research focus: implications of 'Al' for law and the rule of law

2015

- Living with systems that anticipate us
- Mindless agency (ChatGPT avant la lettre)
- Big data spaces (*EU strategy avant la lettre*)
- How does it affect our shared world?
 - and the role and the rule of law



- Implications of 'Al' for law and the rule of law
 - Privacy, fairness the usual suspects
 - More important:
 - 4R AI (robust, resilient, reliable, responsible)
 - Involving methodological integrity and key questions such as:
 - how does design and use of AI shift power relationships?
 - notably when deploying 'legal tech':
 - relationship between client & attorney, democratic players, courts and public administration, contracting parties, justice authorities & individual citizens, justice authorities & those wishing to cross the border

nature

View all journals

Search Q

Log in

Explore content >

About the journal ∨

Publish with us ∨

Subscribe

Sign up for alerts \bigcirc

RSS feed

nature > world view > article

WORLD VIEW | 07 July 2020

Don't ask if artificial intelligence is good or fair, ask how it shifts power



Those who could be exploited by AI should be shaping its projects.

<u>Pratyusha Kalluri</u> ⊠

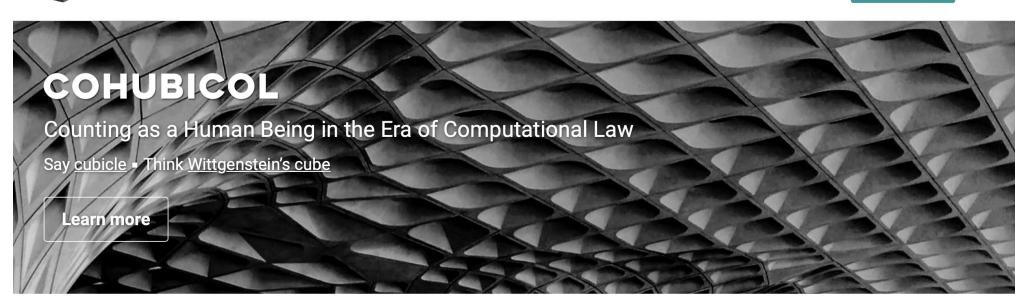
You have full access to this article via Vrije
Universiteit Brussel

Download PDF

Related Articles

The battle for ethical AI at the world's biggest machine-learning conference





It would be nice if all of the data which sociologists require could be enumerated because then we could run them through IBM machines and draw charts as the economists do.

However, not everything that can be counted counts, and not everything that counts can be counted

- William Cameron, Informal Sociology (1963)

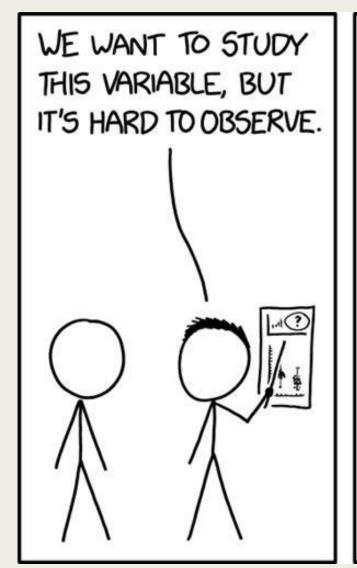


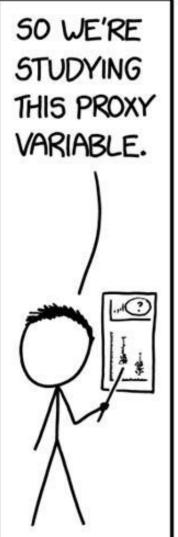


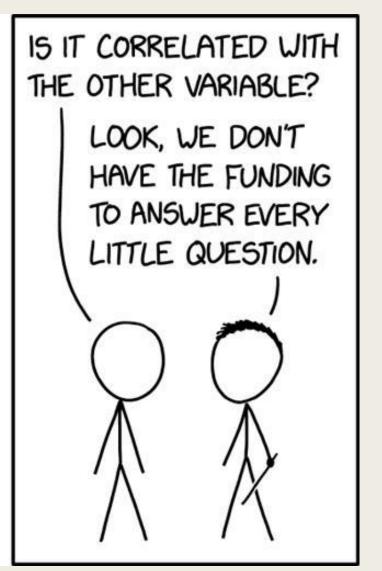
The advent of 'legal tech'

In the context of the ERC ADG we are investigating:

- claims made on behalf of legal technologies
- the substantiation of such claims
 - Mathematical verification, empirical validation, certification
 - Impact on the domain: gaps between requirements and specifications
 - Real-world impact (gap between requirements and real-world goal)







TESTING:

- Accuracy
- **Precision**
- Recall

verification





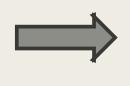


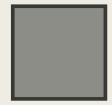
validation

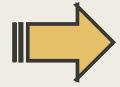












Real World Goals 'do justice'





'select relevant features' 'train an LLM on relevant case



verification



WHEN YOU HEAR THIS:



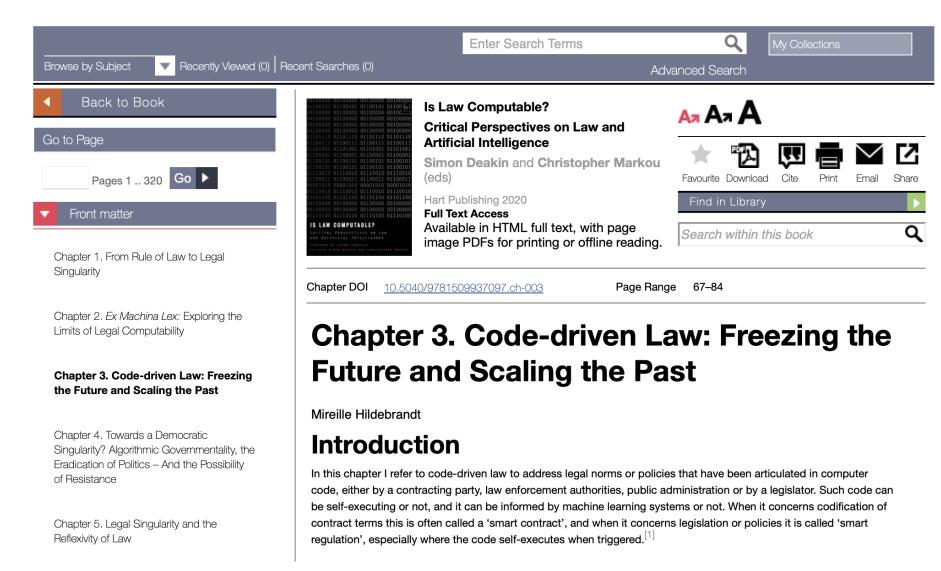
SOFTWARE PROJECT

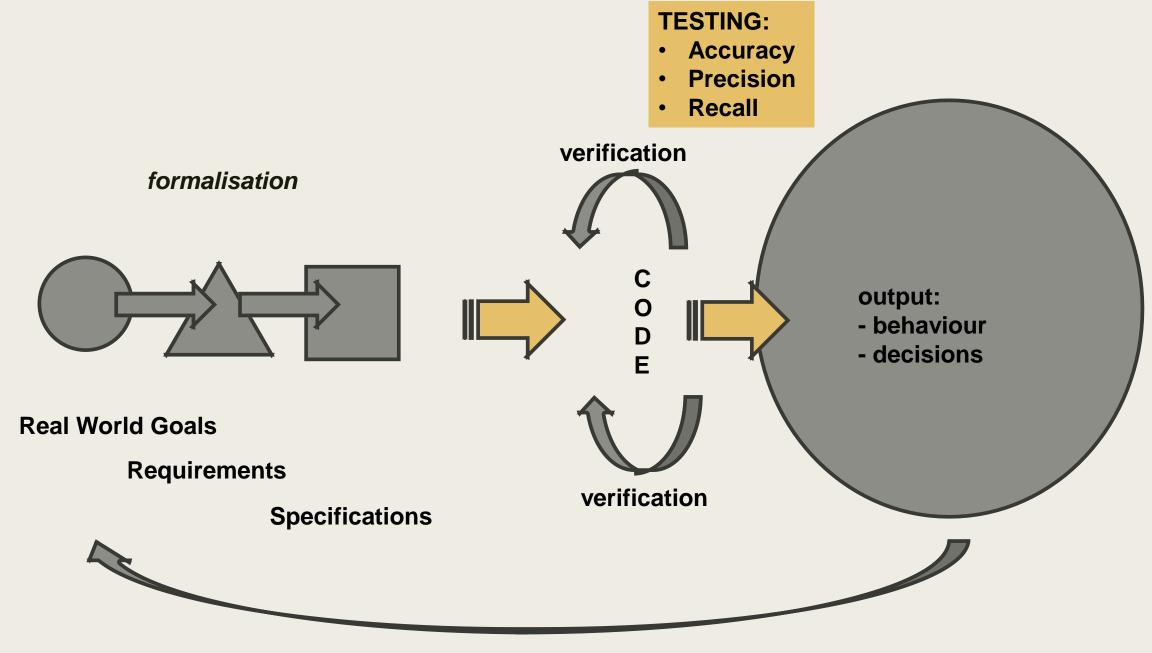
Software, including what some like to call AI, is always running behind.

- Legal expert systems are stuck with the moment they were finalised
- Legal technologies involving ML can only be trained on past data

Prediction is difficult, especially when it's about the future

BLOOMSBURY COLLECTIONS





validation

Key Translations:

- Requirements & Specifications are proxies (formalisation)
- Verification only concerns the internal mathematical check
 - Given the formalisation, is the systems mathematically correct?
- Performance metrics are based on an assumed ground truth
 - Which itself is again a proxy:
 - a training dataset of large legal text corpora (unsupervised)
 - labels that mark features considered relevant (supervised)
- And now we have prompt engineering or RLHF (ChatGPT)
 - Hoping to 'align' the system with our goals



We gratefully acknowledge support from the Simons Foundation and member institutions



All fields



Help | Advanced Search

コアンV > cs > arXiv:2212.14402

Computer Science > Computation and Language

[Submitted on 29 Dec 2022]

GPT Takes the Bar Exam

Michael Bommarito II, Daniel Martin Katz

Nearly all jurisdictions in the United States require a professional license exam, commonly referred to as "the Bar Exam," as a precondition for law practice. To even sit for the exam, most jurisdictions require that an applicant completes at least seven years of post-secondary education, including three years at an accredited law school. In addition, most test-takers also undergo weeks to months of further, exam-specific preparation. Despite this significant investment of time and capital, approximately one in five test-takers still score under the rate required to pass the exam on their first try. In the face of a complex task that requires such depth of knowledge, what, then, should we expect of the state of the art in "Al?" In this research, we document our experimental evaluation of the performance of OpenAl's `text-davinci-003` model, often-referred to as GPT-3.5, on the multistate multiple choice (MBE) section of the exam. While we find no benefit in fine-tuning over GPT-3.5's zero-shot performance at the scale of our training data, we do find that hyperparameter optimization and prompt engineering positively impacted GPT-3.5's zero-shot performance. For best prompt and parameters, GPT-3.5 achieves a headline correct rate of 50.3% on a complete NCBE MBE practice exam, significantly in excess of the 25% baseline guessing rate, and performs at a passing rate for both Evidence and Torts. GPT-3.5's ranking of responses is also highly-correlated with correctness; its top two and top three choices are correct 71% and 88% of the time, respectively, indicating very strong non-entailment performance. While our ability to interpret these results is limited by nascent scientific understanding of LLMs and the proprietary nature of GPT, we believe that these results strongly suggest that an LLM will pass the MBE component of the Bar Exam in the near future.

Comments: Additional material available online at this https URL

Computation and Language (cs.CL); Artificial Intelligence (cs.AI); Machine Learning (cs.LG) Subjects:

arXiv:2212.14402 [cs.CL] Cite as:

> (or arXiv:2212.14402v1 [cs.CL] for this version) https://doi.org/10.48550/arXiv.2212.14402

Submission history

From: Michael Bommarito Ii [view email] [v1] Thu, 29 Dec 2022 18:19:43 UTC (125 KB)

Download:

- PDF
- Other formats



Current browse context: cs.CL

< prev | next > new | recent | 2212

Change to browse by:

CS cs.Al cs.LG

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

Export Bibtex Citation

Bookmark





GPT-4 Passes the Bar Exam

35 Pages • Posted: 15 Mar 2023 • Last revised: 5 Apr 2023

Daniel Martin Katz

Illinois Tech - Chicago Kent College of Law; Bucerius Center for Legal Technology & Data Science; Stanford CodeX - The Center for Legal Informatics; 273 Ventures

Michael James Bommarito

273 Ventures; Licensio, LLC; Stanford Center for Legal Informatics; Michigan State College of Law; Bommarito Consulting, LLC

Shang Gao

Casetext

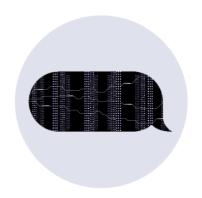
Pablo Arredondo

Casetext; Stanford CodeX

Date Written: March 15, 2023

Abstract

In this paper, we experimentally evaluate the zero-shot performance of a preliminary version of GPT-4 against prior generations of GPT on the entire Uniform Bar Examination (UBE), including not only the multiple-choice Multistate Bar Examination (MBE), but also the open-ended Multistate Essay Exam (MEE) and Multistate Performance Test (MPT) components. On the MBE, GPT-4 significantly outperforms both human test-takers and prior models, demonstrating a 26% increase over ChatGPT and beating humans in five of seven subject areas. On the MEE and MPT, which have not previously been evaluated by scholars, GPT-4 scores an average of 4.2/6.0 as compared to much lower scores for ChatGPT. Graded across the UBE components, in the manner in which a human tast-taker would be, GPT-4 scores approximately 297 points, significantly in excess of the passing threshold for all UBE jurisdictions. These findings document not just the rapid and remarkable advance of large language model performance generally, but also the potential for such models to support the delivery of legal services in society.



₩P EXCLUSIVE

Inside the secret list of websites that make AI like ChatGPT sound smart

By <u>Kevin Schaul</u>, <u>Szu Yu Chen</u> and <u>Nitasha Tiku</u> April 19 at 6:00 a.m.



AI chatbots have exploded in popularity over the past four months,

Inside the Black Box

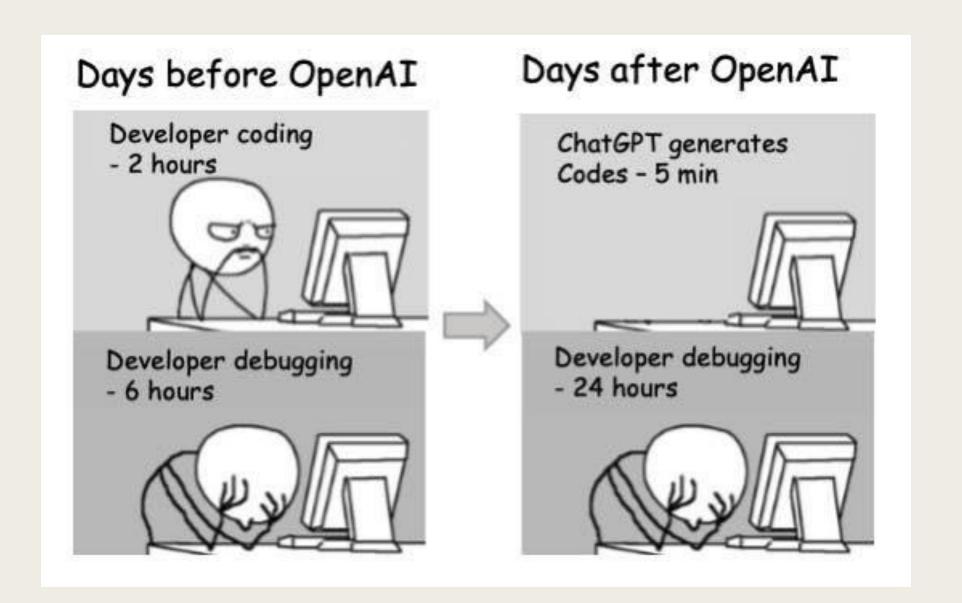
Millions of websites are used to train

Al's biggest chatbots

Home & Garden

To look inside this black box, we analyzed <u>Google's C4 data set</u>, a massive snapshot of the contents of 15 million websites that have been used to instruct some high-profile English-language AIs, called large language models, including Google's T5 and Facebook's LLaMA. (OpenAI does not disclose what datasets it uses to train the models backing its popular chatbot, ChatGPT)

The Post worked with researchers at the Allen Institute for AI on this investigation and categorized the websites using data from Similarweb, a web analytics company. About a third of the websites could not be



What's next?

- Cross-disciplinary perspectives on computational 'law'
- Typology: objectives
- Typology: demonstration
- Typology: a method, a mindset beyond legal technologies

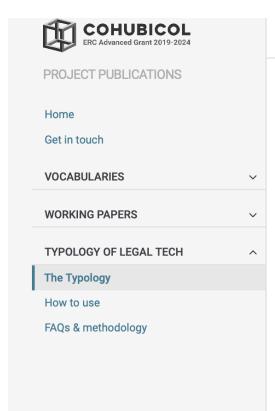
Typology: objectives

- To enable further research into legal technologies, based on our investigation of the substantiation of claims made by their providers and the potential legal impact of their deployment.
- To offer a strategy for review or evaluation of the different types of legal tech.
- To provide a means of comparing aspects of legal tech, especially how they operate at the 'back-end'.
- To make sure our audience (primarily lawyers and computer scientists) can both navigate and understand the information we offer.

What's next?

- Cross-disciplinary perspectives on computational 'law'
- Typology: objectives
- Typology: demonstration
- Typology: a method, a mindset beyond legal technologies

L. Diver, P. McBride, M. Medvedeva, A. Banerjee, E. D'hondt, T. Duarte, D. Dushi, G. Gori, E. van den Hoven, P. Meessen, M. Hildebrandt, 'Typology of Legal Technologies' (COHUBICOL, 2022), available at https://publications.cohubicol.com/typology

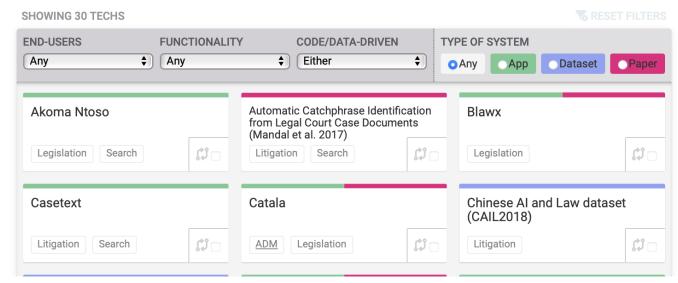


Typology of Legal Technologies

A Method – A Mindset

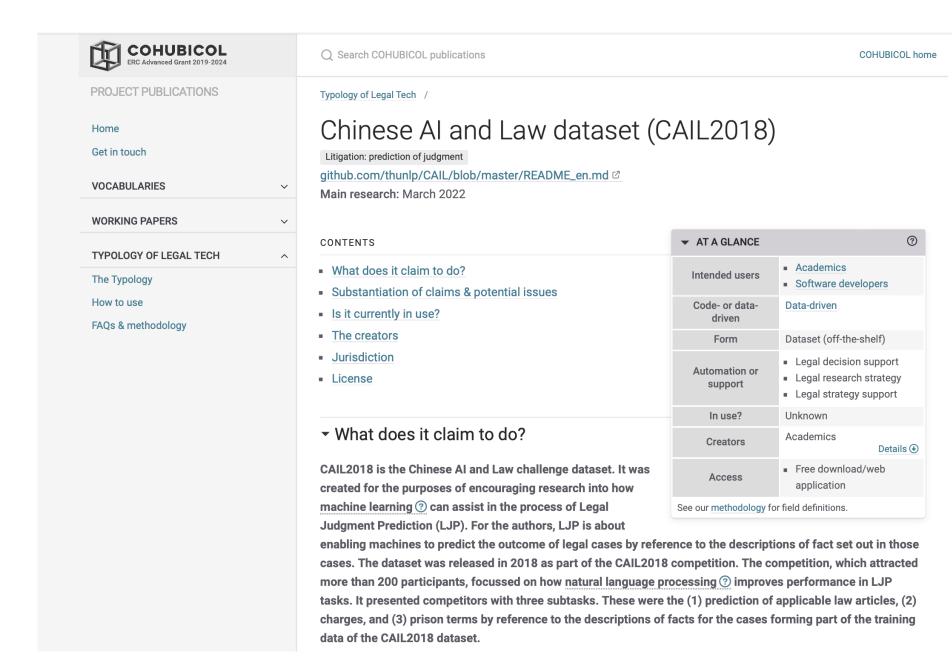
The Typology is a curated set of legal technologies (applications, scientific papers, and datasets) that we handpicked to demonstrate the potential impact on *legal effect* of different types of 'legal tech'. To understand how and why we created this, see the **FAQs & methodology** page.

- Use the filters below to find legal techs you are interested in. Click a system to view its full profile.
- Compare systems by clicking 🗘 □ on one or more systems (view the comparison at the bottom of this page).



Why include datasets?

- Training data sets often stand for a ground truth:
 - 'ground truth' concerns real world issues:
 it cannot be completely and finally computed/formalised
 - meaning that it can be computed/formalised but in different ways
 - And that difference matters



What does it claim to do?

Claimed essential features

- Create a large-scale dataset containing processed data of China Judgments Online, an online repository established by the Supreme People's Court of China.
- Provide a dataset of charges, law articles and prison terms used in Chinese criminal cases.
- **▶ RELEVANT QUOTES**

Claimed rationale and benefits

- To facilitate further research in the field of legal judgment prediction.
- **▶** RELEVANT QUOTES

Claimed design choices

- Each datapoint consists of a case description and three target attributes (labels) the law article cited, charges, and the prison term. The three target attributes correspond to the three subtasks in the CAIL competition. The target attributes are extracted from the original case description using regular expressions ?.
- Law article prediction and charge prediction are framed as text classification tasks, prison term prediction is framed as a regression task in the CAIL competition.
- Only criminal cases were selected from China Judgments Online.
- The cases that would have very infrequent charge or law articles labels are filtered out.
- Cases with multiple defendants were also filtered out to reduce the complexity of the LJP task.
- The dataset includes the fact description (used as input in the LJP task) and the target attributes namely applicable law articles, charges, and prison terms.

Substantiation of claims & potential issues

The dataset is described in two papers (Xiao, C. et al, 2018; Zhong, H. et al, 2018) and on the Github page for the 2018 Chinese Al and Law Challenge Competition 2, where the dataset can be downloaded. A preview of the dataset is available on Hugging Face 2.

Data

- The dataset consists of data collected from China Judgments Online ②, published by the Supreme People's Court of China.
- The time span of the data is not specified.
- The data are stored in a JSON dataset format.
- A preview is available on Hugging Face
 [□] (archived Feb '22 [□]).
- The full dataset is available on Github ☑ (archived Feb '22 ☑).
- "There are two parts of our dataset called CAIL2018-Small and CAIL2018-Large." (Chinese AI and Law Challenge Competition ☑; archived Feb '22 ☑), that contain 196,000 and 1.5 million cases respectively.

Dataset construction

The authors provide some information about how the dataset was constructed. However, no information is provided about how the data was collected (whether, for example, it was scraped from China Judgments Online or downloaded in batch). No information is provided about whether, and if so how, the data was cleaned. The authors provide no information about the completeness of China Judgments Online as a data source.

The dataset has been constructed as follows:

- 1 **5,730,302 criminal documents** were collected from Chinese judgments.
- ² The data is **filtered on 'judgment' documents**, using available metadata.
- 3 The data was filtered to remove cases with more than a single defendant; cases "with those charges and law articles whose frequency is smaller than 30"; and law articles and charges associated with the "top 102 law articles" in Chinese criminal law. (Xiao, C. et al, 2018)
- 4 The target attributes (law articles, charges and prison terms) are constructed using regular expressions ② on the text. It is not known if there is a quality assessment step in case of contradictory candidates or if these data samples were automatically excluded.

Attributes

The attributes of the dataset, along with a short textual description, are set out in Figure 1 below.

- · fact: The description of fact.
- · meta: The label information which contains:
 - o criminals: The defendant in the cases. (There will only be one defendant in the case.)
 - o punish_of_money: The punishing of money in unit RMB.
 - accusation: The defendant's charges.
 - relevant_articles: The relevant articles to the case.
 - o term_of_imprisonment: The term of imprisonment of the defendant. There three more fields in this part:
 - death_penalty: Whether the defendant suffers the death penalty.
 - life_imprisonment: Whether the defendant suffers the life imprisonment.
 - imprisonment: The length of the term of imprisonment in terms of months.

Figure 1: the attributes of the CAIL2018 dataset (Chinese AI and Law Challenge Competition 🖸 ; archived Feb '22 🗹)

An example of the data is shown in Figure 2 below.

```
[2]

"fact": "2015年11月5日上午,被告人胡某在平湖市乍浦镇的嘉兴市多凌金牛制衣有限公司车间内,与被害人孙某因工作琐事发生口角,后被告人树采用
"meta":

{

    "relevant_articles": [234],
    "accusation": ["故意伤害"],
    "criminals": ["胡某"],
    "term_of_imprisonment":
    {
        "death_penalty": false,
        "imprisonment": 12,
        "life_imprisonment": false
    }
}
```

Figure 2: an example of the data (Chinese AI and Law Challenge Competition ♂; archived Feb '22 ♂)

The authors also provide an example in tabular form (Figure 3):

-	Fact	Relevant Law Article	Charge	Prison Term	Defendant
_	被告人胡某	刑法第234条	故意伤害	12个月	胡某
	The Defendant Hu	234th article of criminal law	intentional injury	12 months	Miss /Mr Hn

Judgment prediction

The dataset is used for a Chinese AI and Law Competition in **predicting charges**, **relevant articles** and **term of penalty**.

POTENTIAL TECHNICAL ISSUES

- The examples of the data don't show a specific focus on the time period in which the judgment is made. This
 suggests that any system used to make predictions using this dataset cannot take into account that the laws
 and interpretations of law change over time.
- The original documents already contain the information about the labels, so it is not clear how predicting those labels is helpful for a legal professional.
- The authors do not provide an explanantion of how this experiment could be used to predict actual decisions that will be made by the Chinese courts in the future.
- Court judgments are generally compiled after the decision has been made, therefore the facts of the case are not necessarily representative of the description of the facts prior to the final judgment.
- The authors do not provide any data to be able to predict decisions of the court that have not been made yet.

Rationale and benefits

POTENTIAL TECHNICAL ISSUES

Given the data used for this text classification task it is clear that the system is unable to actually predict
future cases. The authors present a dataset of facts from already made judgments. In order to actually
forecast future decisions of the court the system would require data that was available before the 'predicted'
judgment was made (e.g. case law from a lower court).

References

- Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H. and Xu, J., 2018. Cail2018: A large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478.
- Thong, H., Xiao, C., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H. and Xu, J., 2018. Overview of

Home > Artificial Intelligence and Law > Article

Original Research | Open Access | Published: 25 January 2022

Rethinking the field of automatic prediction of court decisions

Masha Medvedeva , Martijn Wieling & Michel Vols

Artificial Intelligence and Law 31, 195–212 (2023) Cite this article

7896 Accesses 4 Citations 17 Altmetric Metrics

Abstract

In this paper, we discuss previous research in automatic prediction of court decisions. We define the difference between outcome identification, outcome-based judgement categorisation and outcome forecasting, and review how various studies fall into these categories. We discuss how important it is to understand the legal data that one works with in order to determine which task can be performed. Finally, we reflect on the needs of the legal discipline regarding the analysis of court judgements.

Legal effect

- Our focus is on <u>legal effect</u>, that is the effects of written and oral speech acts recognised by law
 - e.g. a civil servant pronouncing a marriage, two parties agreeing to a contract, or a judgehanding down a written judgment
- Legal effect (as we know it) relies on text as its underlying technology
 - any transition in legal practice toward systems that rely on code and data
 - may disrupt the nature and the operation of legal effect.
- Such disruption may affect legal effect and thus legal protection,
 - in order to assess this, the effects must be investigated and anticipated.
- This means considering
 - how legal technologies are and might foreseeably be deployed:
 - by whom, in what contexts, and for what purposes
 - including in ways not intended by the system's provider.
- We summarise this assessment in each Typology profile under the heading *Potential legal impact*.

POTENTIAL LEGAL IMPACT

- Much research in the field of 'legal judgment prediction' does not tackle prediction (in the sense of forecasting) at all. The CAIL2018 dataset does not offer data which enables the prediction of court decisions that have not yet been made. The term 'prediction' may mislead lawyers and policymakers into thinking the field of forecasting judgments is more advanced than it in fact is.
- The original documents already contain information about the labels (legal norms cited, charges, and prison term), so the value to legal practitioners of predicting those existing labels is not evident.
- The descriptions of the facts come from the court judgments, which may not be representative of the facts as set out prior to judgment. They may therefore be an incomplete or partial account of what actually happened.
- The dataset does not include the time period in which the judgments were made, suggesting that predictions
 made using this dataset cannot take into account that <u>legal norms</u> and their <u>interpretations</u> change over time.

What's next?

- Cross-disciplinary perspectives on computational 'law'
- Typology: objectives
- Typology: demonstration
- Typology: a method, a mindset

A Method A Mindset

- ! Avoid the PR!
- Collaborate with CS folk who are capable of
 - Identifying systems that are relevant and reliable
 - Developing an internal critique of CS, while respecting CS methodology
 - Fostering a genuine interest in the law
- In law the point is not to get the outcome right
 - But to get the outcome right for the right reasons
 - Judgment in law is about getting things right in the case at hand
 - It's about precision not accuracy (in case of machine learning)
 - Rules cannot interpret themselves: interpretation is a normative decision

About

News

Subscribe

Team

Blog

CRCL 2023: Computational 'law' on edge

CROSS-DISCIPLINARY RESEARCH IN COMPUTATIONAL LAW

COMPUTATIONAL 'LAW' ON EDGE

The 2nd international conference organised by **COHUBICOL** in collaboration with **CRCL**

General Co-Chairs:

<u>Katie Atkinson</u>, <u>Mireille Hildebrandt</u>, <u>Frank Pasquale</u>, Laurence Diver

20 - 21 November 2023 in Brussels

Hybrid

Registration details to come

Submit extended abstracts at HotCRP 🔀

(Please check the details below before submitting)

Programme committee

Format

Call for abstracts

Conference ethos

Deadlines

Reviewers



m.E.menair