Aligning Learning and Reasoning for Responsible Hybrid Intelligence

Cor Steging [1] Silja Renooij [2] Bart Verheij [1]

[1]Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen [2]Department of Information and Computing Sciences, Utrecht University

Knowledge, Reasoning and Data

A core puzzle in today's artificial intelligence is how knowledge, reasoning and data are connected. To what extent can knowledge used in reasoning be recovered from data with implicit structure? Can such knowledge be correctly recovered from data? To what extent does knowledge determine the structure of data that results from reasoning with the knowledge? Can knowledge be selected such that the output data generated by reasoning with the knowledge has desired properties? By investigating the relations between knowledge, reasoning and data, we aim to develop mechanisms for the verification and evaluation of hybrid systems that combine manual knowledge-based design and learning from data.

Making the Right Decisions for the Wrong Reasons

Many of the state-of-the-art approaches in AI are black box machine learning models. These models learn to perform tasks by exposing them to examples, but their exact internal reasoning often remains unknown. The sub-field of explainable AI aims to solve this issue, by providing explanations to decisions made by black-box models. One key observation in this type of research, is that data-driven models often make the right decision, but do so for the wrong reason. For example, a husky was classified as a wolf because of the snow in the background. This type of unsound reasoning can lead to unwanted, irresponsible behavior. In order to create responsible AI models that behave as intended, we therefore need methods to evaluate and potentially improve their decision making.

Resources for Aligning Learning and Reasoning

To investigate data-driven decision making and the relationship between learning and reasoning, we created a set of resources [5]. Each resources describes a legal domain in the form of a knowledge structure. Additionally, using these knowledge structures, we generate artificial datasets that can be used in experiments. The knowledge structures and their datasets are publicly available.

- In the fictional **Welfare Benefit domain** pensioners may be eligible for a welfare benefit if they satisfy six conditions. These conditions are expressed as logical rules, ranging from simple 1-variable Boolean conditions to more complex multi-variable numerical expressions.
- The **Tort Law domain** is based on real-life Dutch tort law (Articles 6:162 and 6:163 of the Dutch civil code), and describes whether there is a duty to repair damages. These articles are expressed as an argumentative model containing elementary propositions and their arguments and attacks.

Method for Rationale Evaluation and Improvement

To investigate the decision-making of data-driven systems, and evaluate whether they make the right decisions for the right reasons, we have developed a hybrid, model-agnostic method for evaluating and improving rationales [2]:

1. **Measure the performance** of the trained system using contemporary evaluative measures, and proceed if it is sufficiently high;
2. **Design rationale evaluation test sets** for rationale evaluation, targeting selected rationale elements based on expert knowledge of the domain;
3. **Evaluate the rationale** through the performance of the trained system on these rationale evaluation test sets;
4. **Improve the rationale** if needed, by re-training the system on a tailored training dataset, designed using the results from the rationale evaluation.

Evaluating Rationales

We applied our method to each of the resources that we have developed [1]. That means that we train machine learning models on the datasets of the resources and evaluated their performance (step 1), designed rationale evaluation test sets based on the knowledge structures that defined the datasets (step 2), and evaluated the rationale of these machine learning models using the rationale evaluation test sets (step 3).

Results show that our models will generally achieve a high performance (accuracy, F1-score and MCC) in step 1, but lower scores on the rationale evaluation test sets. This implies that the systems make the right decisions, but for the wrong reason. Because our knowledge of the domains in our experiments is exhaustive, we are able to exhaustively test the rationale in a quantitative manner.

Improving Rationales

Based on the results of our rationale evaluation combined with our knowledge of the domain, we were able to pin-point where the machine learning models make mistakes in their decision-making. Using that knowledge, we create new tailored training datasets and repeat the experiment.

Models trained on these tailored datasets have an improved rationale, and therefore make more decisions using the right reasons. Our method can thus be used to not only evaluate, but also improve the decision-making of data-driven AI models.

Rationale Discovery and Explainable AI

In a follow-up experiment, we compared our method for rationale evaluation to two of the most commonly used explainable AI (XAI) techniques: SHAP and LIME. We apply these XAI techniques to the same machine learning models that we applied our own rationale evaluation method on: the models trained on our resources. We discovered that the XAI methods would yield high impact values to all of the relevant features, suggesting a sound decision-making process. Our method for rationale evaluation, however, showed that the decision-making of these models was not sound. These explainable AI methods therefore cannot guarantee a sound rationale. Our hypothesis is that the models used the right features, but in the wrong way. Systems can then make the right decisions, even using the right features, but for the wrong reasons. [3]

Taking the Law more seriously in AI & Law research

When designing data-driven AI models, one should take into account the characteristics of the domain. For example, when it comes to the legal domain, one should take the effects of time into account. It would therefore be unreasonable to test on cases from the past while training on cases from the future. We investigated the effect of some of the design choices in court case prediction research, in terms of performance and the extent to which these are legally reasonable [4]. We studied the choice of performance metric; the effect of including different parts of the legal case; the effect of a more or less specialized legal focus; and the temporal effects of the available past legal decisions.

References

[1] C. Steging, S. Renooij, and B. Verheij.
Discovering the rationale of decisions: Experiments on aligning learning and reasoning.
In *4th EXplainable AI in Law Workshop (XAILA 2021)*, pages 235–239. ACM, 2021.

[2] C. Steging, S. Renooij, and B. Verheij.
Discovering the rationale of decisions: towards a method for aligning learning and reasoning.
In Juliano Maranhão and Adam Zachary Wyner, editors, *ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21 - 25, 2021*, pages 235–239. ACM, 2021.

[3] C. Steging, S. Renooij, and B. Verheij.
Rationale discovery and explainable AI.
In Erich Schweighofer, editor, *Legal Knowledge and Information Systems - JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021*, volume 346 of *Frontiers in Artificial Intelligence and Applications*, pages 225–234. IOS Press, 2021.

[4] C. Steging, S. Renooij, and B. Verheij.
Taking the law more seriously by investigating design choices in machine learning prediction research.
In Mumford J. Odekerken D. Westermann H Lagioia, F., editor, *Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023), June 23, 2023, Braga, Portugal*, pages 49–59. CEUR-WS, 2023.

[5] Cor Steging, Silja Renooij, Bart Verheij, and Trevor Bench-Capon.
Arguments, rules and cases in law: Resources for aligning learning and reasoning in structured domains.
Argument & Computation, 14:235–243, 2023.