Benzmüller, Christoph

Reasonable, Trusted AI through Symbolic Ethico-legal Control and Reflection? (Keynote Abstract)

Date of secondary publication: 05.06.2023

Version of Record (Published Version), Conferenceobject

Persistent identifier: urn:nbn:de:bvb:473-irb-596715

Primary publication

Benzmüller, Christoph: Reasonable, Trusted AI through Symbolic Ethico-legal Control and Reflection? (Keynote Abstract). In: Proceedings of FLAIRS-36, May 14-17, 2023, Clearwater Beach, FL : Florida: LibraryPress@UF, 2023, 2 pp - DOI: 10.32473/flairs.36

Invited Talk Abstracts

General Conference Invited Talks

Knowledge Graphs and the Evolving AI Landscape

Deborah L. McGuinness

Tetherless World Senior Constellation Chair Professor of Computer
Cognitive Science Director Rensselaer Web Science Research Center
Rensselaer Polytechnic Institute

The Artificial Intelligence landscape is changing at an unprecedented pace. Powerful AI tools and services have amazed both the general public as well as many seasoned AI researchers. Like all technologies, however, challenges remain. Many remaining challenges for large language models and generative AI align with strengths of knowledge graphs and semantic AI. In this talk, we will discuss topics including context, abstraction, and provenance. We hope to provide some useful directions for knowledge graph research and applications in today's evolving landscape.

Reasonable, Trusted AI through Symbolic Ethico-legal Control and Reflection?

Christoph Benzmüller

Professor, Chair for AI Systems Engineering, Otto-Friedrich-Universität Bamberg
Professor, Math and Computer Science, Frei University, Berlin, Germany

In formal methods, symbolic specification and verification are prominent and successful means of achieving trust and security in software and hardware development. Due to their opaque, statistical nature, combined with their dependence on typically imperfect data, modern subsymbolic AI models can be seen as antipodes to systems developed according to this formal methods paradigm. In subsymbolic AI, the sharp concept of 'trust' has thus been replaced by the fuzzy and inconclusive concept of 'trustworthiness', which seems to be insufficient for most critical applications. I will argue that appropriate hybrid (or neuro-symbolic) AI architectures are a promising option for overcoming this dichotomy. Not only could they reintroduce a sharper notion of trust, but they could enable run-time alignment between socially legitimated, regulative ethico-legal theories and trained models. The key idea is to focus less on explaining the imperfection of trained models, but rather on the independent evaluation and justification (with symbolic means) of their proposed actions in a given application context. Adopting this perspective, I will outline recent collaborative research on (i) ethico-legal control and reflection architectures, (ii) logico-pluralistic normative reasoning using the meta-logical knowledge representation and reasoning methodology LogiKEy, and (iii) recent progress in the

use of higher-order interactive and automated theorem provers to support the automation of (not only) normative reasoning in the LogiKEy framework.

Suggested Readings

- Designing Normative Theories for Ethical and Legal Reasoning: LogiKEy Framework, Methodology, and Tool Support. *Artificial Intelligence*, 2020. http://doi.org/10.1016/j.artint.2020.103348
- Reasonable Machines: A Research Manifesto. *KI 2020,* Springer, 2020. http://doi.org/10.1007/978-3-030-58285-2_20
- Universal (Meta-)Logical Reasoning: Recent Successes. *Science of Computer Programming*, 2019. https://doi.org/10.1016/j.scico.2018.10.008
- Who Finds the Short Proof? *Logic Journal of the IGPL*, 2023. http://doi.org/10.1093/jigpal/jzac082

AI and NLP Challenges, Solutions, and Gaps in the age of Chat GPT

Bonnie J. Dorr
Professor of Computer Science
the University of Florida

This talk presents challenges, solutions, and gaps in AI and Natural Language Processing (NLP), with an emphasis on the need for explainability in the era of ChatGPT. Examples include: machine translation of human languages, ask detection for defending against social engineering attacks, and stance detection for extracting attitudes from social media. Past, current, and future projects face several challenges: (a) brittleness of rule-based linguistic principles for large-scale processing; (b) shallowness of statistical methods and neural language models for understanding implicit information; and (c) lack of “explainability” amidst ever-increasing numbers of black-box models. A case is made for hybrid approaches that combine linguistic generalizations with statistical and neural models to handle implicitly conveyed information (e.g., beliefs and intentions), and also for the implementation of an “explainable” propositional representation that supports the ability of developers and end users to understand what is going on inside the AI system. Questions of interest range from “What is the social engineer’s underlying goal in a two-way interaction?” to “What beliefs support individuals’ attitudes regarding pandemic interventions?” to “How does targeted influence impact attitudes online?”. Such information is generally not extractable from large language models alone and, moreover, such models are hampered in that they are too large to retrain on a regular basis by the average researcher, developer, or customer. Representative examples of ChatGPT output are provided to illustrate areas where more exploration is needed, particularly with respect to extraction of intentions and task-specific goals.

Special Track Invited Talks

Special Track: Applied Natural Language Processing

Knowledge Discovery in Aircraft Maintenance Records

Nobal Niraula
Boeing Research & Technology

Aircraft maintenance records log day-to-day maintenance activities performed on in-service aircraft and possess a wealth of crucial knowledge related to parts and systems such as part names and associated issues. Mining such knowledge is critical for prognostics and health management, and in particular is essential to improve safety and quality, lower lifecycle maintenance cost, minimize downtime, improve parts inventory related to an aircraft, and improve manufacturing quality and rework. The maintenance records conveyed in unstructured text are very noisy with frequent use of local jargon, ad-hoc acronyms, misspellings, and non-standard abbreviations. In addition, general methods for preprocessing and knowledge extraction do not work effectively for domain-specific cases. Thus, automatically extracting knowledge from maintenance records is a daunting task. This presentation covers the practical constraints and challenges in discovering knowledge from aircraft maintenance records. It also covers the practical Natural Language Processing and Machine Learning methods to address the challenges.