Legal, Ethical, Customizable Artificial Intelligence

Joseph A. Blass
McCormick School of Engineering and Pritzker School of Law, Northwestern University
joeblass@u.northwestern.edu

Abstract

To be effective, useful, safe, and legal, AI must obey the laws of its users' societies and (where legal) its users' ethical intuitions. But laws and ethics can be difficult for people to express. My research involves ethical and legal instruction by example: synthesizing cases, applying synthesized principles, and explaining those applications.

Challenge and Research Goals

Artificial Intelligence (AI) systems now make independent decisions with legal and ethical consequences, and will make many more in the future. Self-driving cars, AI medical, legal and financial systems, and others can take actions with significant consequences for life, health, and wealth. AI systems must obey the laws of the societies within which they operate. But the law is silent on many moral and ethical principles: these should not be imposed by the AI or its creator upon the user, but should be personalizable. Additionally, while some statutory law may be straightforwardly representable as rules for an AI system to follow, common law (derived from judicial interpretation) rarely is; similarly, end users may struggle to express their ethical norms in a way that leads the AI to learn that which it is meant to. However, common law and ethical principles are naturally embodied in the cases that apply them.

In my research, I teach an AI system ethical and legal principles using descriptive, structured examples. The system extracts the legal or ethical principles underlying those examples and applies those principles to new cases in a transparent, explainable way. As a member of a joint JD/PhD program, my research focuses not only on legal instruction, but analyzes legal reasoning and accountability by AI systems to ensure their legal behavior.

Human ethical intuitions vary across and within cultures, and there are a range of distinct ethical principles humans accept each other as holding (Sachdeva et al. 2011). Just as AIs sold in foreign countries must abide by laws there, not only those of their country of manufacture, the creators and programmers of AI systems should not force its users to abide by the creators' ethical principles where the users disagree (for example, in determining when it is appropriate to tell a white lie to spare someone's feelings).

In the United States and many countries, there are two sources of law: statutes passed by legislatures, and common law decisions by judges interpreting and applying the law, using statutes, their precedents and those of the courts above them, and explicit reasoning. Common law decisions (and statutes) do not always define easily interpretable, expressible, and applicable rules. Similarly, ethical norms may sometimes seem easy to express but often admit exceptions. Expressing the norm as a rule, with all its exceptions and how they apply, may be too much for the average user of an AI system. However, like common law decisions applying laws to facts, people can provide examples that illustrate when a rule applies, and when it does not. Both common law rules and ethics thus lend themselves to illustration over description. Through examples of defaults and exceptions, the complexities of rules emerge. My research explores such illustration and synthesis.

Research Tools: MoralDM, SME, and SAGE

My work grew out of MoralDM (Dehghani et al. 2009), a computer model of moral reasoning that takes in natural language moral dilemmas, extracts from it Cyc-derived predicate logic representations using a natural language understanding (NLU) system, and uses first-principles reasoning and the Structure Mapping Engine (SME) over resolved cases to make human-consistent moral decisions.

SME, based on the Structure Mapping Theory of analogy (Gentner 1983), creates and draws inferences from an alignment between two relational cases. Analogy is useful for ethical and legal reasoning because such cases are defined not only by the nature of actions, actors, and events, but crucially by the relationships between these.

The Sequential Analogical Generalization Engine (SAGE) uses SME to build case generalizations, emphasizing shared structures and facts and deprecating case-

specific ones. SAGE uses a case library of examples and generalizations. Generalizations contain facts from constituent cases: abstract entities replace non-identical corresponding ones; probabilities represent the proportion of cases each fact is in. Given a probe, SAGE retrieves a similar case from its library. Strongly matching cases are assimilated; weakly matching cases are added as examples.

Achievements

MoralDM originally matched over all resolved dilemmas, which is cognitively implausible and computationally expensive. I extended it to generalize moral principles using SAGE (Blass & Forbus, 2015). Generalizations led to more human-like judgments than using ungeneralized cases.

To explore trade-offs between ethical rules, I used defeasible logic to model ethical norms in virtual characters (Blass & Horswill, 2015). Defeasible logic encodes default rules that can be overcome or traded off, which is necessary for ethical dilemmas or laws that permit exceptions.

The world is complex; human stories assume the receiver understands unstated implications (e.g., pain is bad). Learning from examples requires making such inferences. Analogical Chaining uses analogies over simple 'commonsense' cases to repeatedly enrich a case (Blass & Forbus 2016). The system learns such commonsense interactively through short natural language *microstories*, which were generalized using SAGE (Blass & Forbus 2017). Understanding the nature of the situation at hand is crucial to being able to make proper inferences about it – a system must understand the difference between "Jim threw rocks at Bob" and "Jim threw crumpled paper at Bob" if it is to reason properly about these two descriptively similar cases.

Future Work

In pursuit of a joint JD/PhD degree, I now focus on computational models of legal reasoning. Synthesizing rules from common-law decisions requires reasoning at different levels of representation and abtraction. I want to model such reasoning in a series of experiments using SAGE. I will seek a set of consistent real-life cases illustrating subtly different rules in different jurisdictions. It is unclear the extent to which these cases can be encoded by NLU systems that generate structured, propositional representations. I suspect that legal language is still too complex for such systems, but will quantify the extent to which such a task is achievable, and how much remains to be done. To the extent that the NLU system cannot translate the cases from English to CycL, they will be translated by experts.

Legal precedents provide a rich experimental playground, insofar as an AI's reasoning based on a series of precedents can be tested by comparing its induced rule to that of the next real-life case. The first experiment will simply be to quantify the extent to which SAGE can synthesize, from precedent cases, the same rules that judges apply in subsequent decisions, to identify the limit points of the technology, and to see what changes can be made manually to lead to the right results.

Subsequent experiments will seek to automate those fruitful manual processes. The second planned experiment will be to explore rerepresentation in the construction of generalizations and mappings, in order to improve rule synthesis and application. Legal reasoning proceeds deliberately, beginning with the most salient elements of a case: the third experiment will be to explore attention and focus mechanisms to learn what facts are important (or worthy of rerepresentation) in generalization construction and mapping (i.e., which must play important roles in mappings and inferences), and which should be deprecated. A possible final experiment will involve exploring how to convert generalized structures into explicit horn-clause like rules that can be followed using first-principles reasoning.

My thesis will also have an important theoretically focused review and analysis component examining the current and future limitations and capabilities of computational legal models, with the techniques I now use in my own research, and others. For this component, the goal is to make progress mapping the space at the intersection of Jurisprudence and AI, with a firm footing in both fields.

Combining the theoretical and experimental components of the thesis, my thesis goal is to clearly define what it would take to instruct an AI to behave legally and ethically using example cases and to make significant progress towards the goal of doing so, assessing both its learning and its ability to apply what it has learned.

References

Blass, J. and Forbus, K. 2015. Moral Decision-Making by Analogy: Generalizations vs. Exemplars. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, Austin, TX.

Blass, J. A. and Forbus K. D. 2017. Analogical Chaining with Natural Language Instruction for Commonsense Reasoning. *Procs. Of the 31st AAAI Conference on Artificial Intelligence.* San Francisco, CA. pp. 4357-4363.

Blass, J. A. and Forbus, K. D. 2016. Modeling Commonsense Reasoning via Analogical Chaining: A Preliminary Report. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, Philadelphia, PA, August.

Blass, J.A. and Horswill, I. 2015. Implementing Injunctive Social Norms using Defeasible Reasoning. *Workshop on Social Believability, Procs of the Eleventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Santa Cruz, CA.

Dehghani, M., Sachdeva, S., Ekhtiari, H., Gentner, D., & Forbus, K. (2009). The role of cultural narratives in moral decision making. *Procs of the 31st Annual Conf. of the Cog. Sci. Society*.

Gentner, D. 1983. Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science* 7(2).

Sachdeva, S., Singh, P., and Medin, D. 2011. Culture and the quest for universal principles in moral reasoning. *International Journal of Psychology*, *46*(3)