On Requirements and Design Criteria for Explainability in Legal AI

Martijn VAN OTTERLO [a]Martin ATZMUELLER [a]

[a] *Dept. of Cognitive Science and AI, Tilburg University, The Netherlands;*

Abstract. In this position paper we briefly describe two approaches to explainable artificial intelligence for legal, and other, domains. One aims at designing explainable AI using legal requirements and the other deals with design criteria for computational, ethical reasoning systems in which explainability is a core element.

Ethical challenges of artificial intelligence (AI) are rising as technological advances are widely spread [1,2]. In addition to critiques from legal and sociology scholars who study the influence and regulation of algorithms, nowadays AI researchers themselves get actively involved by *creating* AI technology that is intrinsically *responsible*, *transparent* and especially *explainable* [3]. AI researchers contribute through, for example, *fair machine learning*, *value-based AI* and *ethical reasoning systems*, c. f., [4]

Recently, the concept of transparent and explainable models has gained a strong focus and momentum in the machine learning and data mining community, e. g., [5,6]. Some focus on specific models, e. g., tree-based [7] or pattern-based approaches [8] to better understand where a classifier does not work using local pattern mining techniques. Here, also, methods for associative classification, e. g., class association rules [9] can be applied for obtaining explicative, i. e., transparent, interpretable, and explainable models [10]. Then, individual steps of a classification (i. e., a decision can be traced-back to the model, similar to *reconstructive explanations*, c. f., [11] on several explanation dimensions [5]. For model agnostic explanation, e. g., [12], general directions are given by methods considering counterfactual explanation, e. g., [13,14], data perturbation and randomization techniques as well as interaction analysis methods, e. g., [15].

1. Designing Explainable AI Using Legal Requirements: The KORA Approach

One prominent method for matching legal (normative) requirements with technical requirements and specific implementation options, is the KORA method [16,17]. It has been applied in several contexts, e. g., in an integrated approach for socio-technical design and development of ubiquitous computing applications [18,19].

The basic KORA method aims at acquiring technical implementations based on legal requirements. It is built on a four step process model, starting with legal requirements that are mapped to legal criteria which are then matched with functional requirements. The legal requirements are typically derived from application specific legal provisions, e. g., given by the GDPR. These legal provisions are then made more concrete in the second step, also including technical functions as well as legal and social aspects, formalized in specific criteria. These criteria are then mapped to functional requirements in

Preprint of: Martijn van Otterlo and Martin Atzmueller (2018) On Requirements and Design Criteria for
Explainability in Legal AI. In: Proc. Proc. Workshop on Explainable AI in Law (XAILA), co-located with
International Conference on Legal Knowledge and Information Systems (JURIX), Groningen, Netherlands

the third step, e. g., supported by domain experts or based on utilizing design patterns. In the final fourth step, these functional requirements then map to specific implementation choices. Relating to common process models in software engineering, the first three steps basically aim at requirements analysis, whereas the last step involves specific methods, patterns, and techniques relating to the concrete instantiations.

For enabling explainable AI systems, KORA provides an effective approach, by matching the (abstract) legal provisions with concrete implementation choices, e. g., by implementing specific explainable models described above. For that, a categorization of these methods according to legal criteria (second step of KORA) is needed, connected to functional requirements. Then, a semi-automatic approach can be provided for mapping these criteria and its functional implementation. Furthermore, design patterns for explainable systems can also be incorporated here, by abstracting from specific design choices to more general classes of explanation patterns that are described in terms of their "explanation criteria" as well by the included "legal criteria".

2. A General Research Strategy for Ethical AI: The IntERMeDIUM approach

As a synthesis of a mix of ideas on learning, ethical codes and intentional agents [20,2] IntERMeDIUM is a research strategy to develop ethical AI systems. IntERMeDIUM refers to Joseph Licklider's connection between humans and "the body of all knowledge", which is increasingly governed by AI in our society c. f., [2]. To unite human and machine ethics, a *code of ethics* can be seen as a *moral contract* between human and machine. The acronym covers the main directions on which to focus research efforts on:

Intentional: The bridge between humans and machines consists of the right ontology of the (physical) world and the right level of description: beliefs, desires, intentions and goals. AI should be understood as rational agents.

Executable: The beliefs and desires of the AI need to be embedded as code that can be executed. Instead of asking code of ethics to be enforceable by punishing bad behavior after the fact, executable codes of ethics are biased by the code to ensure the right ethical behavior. [21](p16): *"Ethics must be made computable in order to make it clear exactly how agents ought to behave in ethical dilemmas".*

Reward-based: AI's ethical reasoning is based on the human values in a particular domain. The core values come from the code of ethics used to bias the agent. In addition, AIs *finetune* their ethical behavior over time by adjusting relative values using data, feedback and experience. Experience from human actors is vital here, since they typically solved ethical dilemmas that arose thus far cf. [2].

Moral: The focus of AI implementations here is on the moral dimension. Other skills will be developed elsewhere, including perception, mobile manipulation, reasoning with uncertainty, language interpretation and more.

Declarative: All ethical bias in the AI is declarative knowledge and can be inspected at all times. Ethical inferences in specific circumstances can be explained in human-understandable terms. [21] (p17): *"What is critical in the explicit ethical agent distinction in our view, lies not only in who is making the ethical judgments (the machine versus the human programmer) but also in the ability to justify ethical judgments that only an explicit representation of ethical principles allows."* Ethical bias and learned ethical knowledge can be shared with other AIs and laws and regulations, such as the GDPR, can be implemented in the declarative bias to further bias the behavior of the AI towards legal compliance.

Inductive: The AI is a *learning* agent. All knowledge that can not be injected as a declarative bias needs to be learned from experience or obtained from other AIs or humans. The AI's knowledge will typically not be complete, and learning should be continuing and life-long. Advanced machine learning needs to be implemented that allows the AI to ask human specialists for advice in various ways.

Utilitarian: AI are utilitarian (collective consequentialist) morally reasoning agents. Protection of the rights of individuals is ensured by demanding that values and decision logic are declarative, open for inspection and transparent.

Machine: The AI is a *machine*. The slow migration from human specialists to AI implementations in any domain requires that we should shift focus from humans to machines for the main operational aspects domains ranging from autonomous cars, libraries, and surely legal practices.

The IntERMeDIUM is a general strategy for ethical, explainable AI systems. We claim that especially in the legal domain declarativeness, explainability and the AI's capability of engaging in a dialogue with humans to discuss and learn ethical and legal norms and values, is vital. Inspiration on how to translate human laws and ethical codes into declarative bias can come from e.g. medical [21] and autonomous driving [22] domains.

A first *instantiation* of IntERMeDIUM are *Declarative decision-theoretic ethical programs* (DDTEPs) [20]. They form a novel way to formalize *value-based ethical decision making* to help building understandable AI systems that are *value aligned* [23] in stochastic domains. The idea is to formalize what is known explicitly in the model, use *learning* to fill in knowledge gaps, and to use *reasoning* to obtain (optimal) decisions. DDTEPs fit into logical approaches for ethical (or: value-driven) reasoning [21] but also *relational reinforcement learning* [24] and provides novel opportunities for *explanation-focused* computations [25]. DDTEPs prove successful for toy ethical domains but could generally be applied to any kind of ethical reasoning where (some) domain knowledge is available, especially legal domains.

3. Next Steps

Both approaches are ways to approach the construction of modern AI systems that could be employed in legal domains in a transparent way. Such work is only yet starting and progress needs to be made by answering a couple of questions first: i) What exactly *is* an explanation in the legal context? ii) What are specific legal requirements and criteria for explainable AI systems? iii) How to map legal criteria and functional criteria to each other for explainable AI? iv) How to abstract functional requirements for legal AI into (legal) design patterns? v) How to formalize existing codes of ethics, legal procedures, legal design patterns and legal practices into AI programs? vi) How can AIs interactively learn and explain their functioning in *human-understandable* terms?

Above all, AI and legal specialists need to engage in a dialogue when tackling explainability questions in legal AI. AI systems need to be explainable in order to be *trustworthy* [3] but AI researchers need to know *which* (types of) explanations are most useful to legal scholars. Much work awaits.

References

[1] B.D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2), 2016.

[2] M. van Otterlo. Gatekeeping Algorithms with Human Ethical Bias: The Ethics of Algorithms in Archives, Libraries and Society, 2018. https://arxiv.org/abs/1801.01705.

[3] S. Bobek G.J. Nalepa, M. van Otterlo and Martin Atzmueller. From context mediation to declarative values and explainability. In *Proc. IJCAI Workshop on Explainable Artificial Intelligence (XAI)*, 2018.

[4] M. van Otterlo. Ethics and the value(s) of artificial intelligence. *Nieuw Archief voor Wiskunde*, 5/19(3):206–209, 2018.

[5] M. Atzmueller and T. Roth-Berghofer. The Mining and Analysis Continuum of Explaining Uncovered. In *Research and Development in Intelligent Systems XXVII*, pages 273–278. Springer, 2011.

[6] O. Biran and C. Cotton. Explanation and Justification in Machine Learning: A Survey. In *IJCAI-17 Workshop on Explainable AI*, 2017.

[7] G. Tolomei, F. Silvestri, A. Haines, and M. Lalmas. Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking. In *Proc. KDD*. ACM, 2017.

[8] W. Duivesteijn and J. Thaele. Understanding Where Your Classifier Does (Not) Work – The SCaPE Model Class for EMM. In *Proc. ICDM*, pages 809–814. IEEE, 2014.

[9] M. Atzmueller, N. Hayat, M. Trojahn, and D. Kroll. Explicative Human Activity Recognition using Adaptive Association Rule-Based Classification. In *Proc. IEEE Future IoT*. IEEE, 2018.

[10] M. Atzmueller. Onto Explicative Data Mining: Exploratory, Interpretable and Explainable Analysis. In *Proc. Dutch-Belgian Database Day*. TU Eindhoven, Netherlands, 2017.

[11] M. R. Wick and W. B. Thompson. Reconstructive Expert System Explanation. *Artificial Intelligence*, 54(1-2):33–70, 1992.

[12] M.T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-Precision Model-Agnostic Explanations. AAAI, 2018.

[13] D. R Mandel. Counterfactual and Causal Explanation: From Early Theoretical Views To New Frontiers. In *The Psychology of Counterfactual Thinking*, pages 23–39. Routledge, 2007.

[14] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. 2017.

[15] A. Henelius, K. Puolamäki, and A. Ukkonen. Interpreting Classifiers through Attribute Interactions in Datasets. *Proc. ICML Workshop on Human Interpretability in Machine Learning*, 2017.

[16] V. Hammer, U. Pordesch, and A. Roßnagel. *Betriebliche Telefon- und ISDN-Anlagen rechtsgemäß gestaltet*. Edition SEL-Stiftung. Springer, 1993.

[17] A. Roßnagel and V. Hammer. KORA. Eine Methode zur Konkretisierung rechtlicher Anforderungen zu technischen Gestaltungsvorschlägen für Informations- und Kommunikationssysteme. *Infotech*, 1:21 ff., 1993.

[18] K Geihs, J.M. Leimeister, A. Roßnagel, and L. Schmidt. On socio-technical enablers for ubiquitous computing applications. In *Proc. Workshop on Enablers for Ubiquitous Computing and Smart Services*, pages 405–408, Izmir, Turkey, Juli 2012. IEEE.

[19] M. Atzmueller, K. Behrenbruch, A. Hoffmann, M. Kibanov, B.-E. Macek, C. Scholz, H. Skistims, M. Söllner, and G. Stumme. *Socio-technical Design of Ubiquitous Computing Systems*, chapter Connect-U: A System for Enhancing Social Networking. 2014.

[20] M. van Otterlo. From Algorithmic Black Boxes to Adaptive White Boxes: Declarative Decision-Theoretic Ethical Programs as Codes of Ethics. In *Proc. AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 2018.

[21] M. Anderson and S.L. Anderson. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, 28:15–26, 2007.

[22] N.J. Goodall. Machine ethics and automated vehicles. In *Road Vehicle Automation, Lecture Notes in Mobility, Springer*, 2014.

[23] J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch. Alignment for Advanced Machine Learning Systems, 2017. MIRI (unpublished) https://intelligence.org/2016/07/27/alignment-machine-learning/.

[24] M. van Otterlo. Solving Relational and First-Order Markov Decision Processes: A Survey. In M.A. Wiering and M. van Otterlo, editors, *Reinforcement Learning: State-of-the-art*, chapter 8, pages 253–292. Springer, 2012.

[25] M. van Otterlo. Intensional Dynamic Programming: A Rosetta Stone for Structured Dynamic Programming. *Journal of Algorithms*, 64:169–191, 2009.