Reasonable Machines: A Research Manifesto

Christoph Benzmüller and Bertram Lomfeld

June 14, 2020

Reasonable Machines: A Research Manifesto

Christoph Benzmüller[1][0000−0002−3392−3093] and Bertram
Lomfeld[2][0000−0002−4163−8364]

[1] Institute of Computer Science, Freie Universität Berlin, Berlin, Germany
c.benzmueller@fu-berlin.de
[2] Department of Law, Freie Universität Berlin, Berlin, Germany
bertram.lomfeld@fu-berlin.de

Abstract. Future intelligent autonomous systems (IAS) are inevitably deciding on moral and legal questions, e.g. in self-driving cars, health care or human-machine collaboration. As decision processes in most modern sub-symbolic IAS are hidden, the simple political plea for transparency, accountability and governance falls short. A sound ecosystem of trust requires ways for IAS to autonomously justify their actions, that is, to learn giving and taking reasons for their decisions. Building on social reasoning models in moral psychology and legal philosophy such an idea of »Reasonable Machines« requires novel, hybrid reasoning tools, ethico-legal ontologies and associated argumentation technology. Enabling machines to normative communication creates trust and opens new dimensions of AI application and human-machine interaction.

Keywords: Trustworthy and Explainable AI · Ethico-Legal Governors · Social Reasoning Model · Pluralistic, Expressive Normative Reasoning

1 Introduction

Intelligent autonomous systems (IASs) are rapidly entering applications in industry, military, finance, governance, administration, healthcare, etc., leading to an historical transition period with unprecedented dynamics of innovation and change, and with unpredictable outcomes. Politics, regulatory bodies, indeed society as a whole, are challenged not only with keeping pace with these potentially disruptive developments, but also with staying ahead and wisely guiding the transition. Fostering positive impacts, while preventing negative side effects, is a balanced vision shared within most of the numerous ethical guidelines of the last years on trustworthy AI, including the European Commission's most recent White Paper on AI [5], proposing the creation of an *"ecosystem of excellence"* in combination with an *"ecosystem of trust"*.

We think that real *"Trustworthy AI by Design"* demands IASs, which are able to give and take reasons for their decisions to act. Such »Reasonable Machines« require novel, hybrid reasoning tools, upper ethico-legal ontologies and associated argumentation technology to be utilised in practice for assessing, justifying and controlling (externally and internally) the behaviour of IASs

with respect to explicitly encoded legal and ethical regulation. We envision this technology to be integrated with an on-demand, cloud-based workbench for pluralistic, expressive regulatory reasoning. This would foster knowledge transfer with industry, research, and educational institutions, it would enable access to critical AI infrastructure at scale with little risk and minimal costs, and, in the long run, it could support dynamic adjustments of regulating code for IASs in the cloud via politically and socially legitimated processes.

Paper structure: Section 2 formulates objectives for Reasonable Machines, and section 3 provides models for them building on moral psychology and legal philosophy. Section 4 outlines modular steps for research and implementation of Reasonable Machines; this leverages own prior work such as the LogiKEy methodology and framework for designing normative theories for ethical and legal reasoning [3], which needs to be combined and extended with an upper-level value ontology [16] and further domain-level regulatory theories for the assessment and explanation of ethical and legal conflicts and decisions in IASs.

2   Reasonable Machines: Objectives

The need for some form of "moral machines" [22] is no science fiction scenario at all. With the rise of autonomous systems in all fields of life including highly complex and ethically critical applications like self-driving cars, weapon systems, healthcare assistance in triage and pandemic plans, predictive policing, legal judgement supports or credit scoring tools, involved AI systems are inevitably confronted-with and deciding-on moral and legal questions. One core problem with ethical and legal accountability or even governance of autonomous systems is the hidden decision process (black box) in modern (sub-symbolic) AI technologies, which hinders transparency as well as direct intervention. The simple plea for transparency disregards technological realities or even restrains much needed further developments.[3]

Inspired by moral psychology and cognitive science, we envision the solution in the development of independent, symbolic logic based safety-harnesses in future AI systems [8]. Such "ethico-legal governors" encapsulate and interact with black box AI systems, and they will use symbolic AI techniques in order to search for possible justifications, i.e. reasons, for their decisions and (intended) actions with regard to some formally encoded ethico-legal theories defined by regulating

---

[3] While interpreting, modeling and explaining the inner functioning of black box AI systems is relevant also with respect to our Reasonable Machines vision, such research alone cannot completely solve the trust and control challenge. Sub-symbolic AI black box systems (e.g. neural architectures) are suffering from various issues (including adversarial attacks and influence of bias in data) which cannot be easily eliminated by interpreting, modeling and explaining them. Offline, forensic processes are then required such that the whole enterprise of turning black box AI systems into fully trustworthy AI systems becomes a challenging multi-step engineering process, and such an approach is significantly further complicated when online learning capabilities are additionally foreseen.

bodies. The symbolic justifications computed at this abstract level thus provide a basis for generating explanations about why a decision/action (proposed by an AI black box system) is ethico-legally legitimate and compliant with respect to the encoded ethico-legal regulation.

Such an approach is complementary to, and as an additional measure more promising than, explaining the inner (mis-)functioning of the black box AI system itself. Symbolic justifications in turn enable the development of further means towards a meaningful and robust control and towards human-explanation and -interaction. The REASONABLE MACHINES idea outlines a genuine approach of trustworthiness by design proposing, in psychological terminology [13], a slow, rational (e.g. symbolic) "System 2" layer in responsible IASs to justify and control their fast, *"intuitive"*, but opaque (sub-symbolic), "System 1" layer computations.

REASONABLE MACHINES research aims at analyzing and constructing ways how intelligent machines could socially justify their actions at abstract level, i.e. give and take moral and legal reasons for their decisions to act. Reason is based on reasons. This is true as much for artificial as for human intelligent agents. The "practical reasonableness" of intelligent agents depends on their moral abilities to communicate socially acceptable reasons for their behavior [10]. Thus, the exploration of methods and tools enabling machines to generate normative reasons (which may be independent of underlying black box architectures and opaque algorithms) smoothes the way for more comprehensive artificial moral agency and new dimensions of human-machine communication.

The core objectives of REASONABLE MACHINES technology are:

- enabling argument-based explanations & justifications of IAS decisions,
- enabling ethico-legal reasoning about, and public critic of, IAS decisions,
- facilitating political and legal governance of IAS decision making,
- evolving ethico-legal agency and communicative capacity of IASs,
- enabling trustworthy human-interaction by normative communication,
- fostering development of novel neuro-symbolic AI architectures.

3   Artificial Social Reasoning Model (aSRM)

The black box governance problem has an interesting parallel in human decision making. Most actual models in moral psychology consider emotional intuition to be *the* (or at least one) initial driving force of human action which is only afterwards (or with a second significantly slower system) rationalized with reasons [11, 13]. Within a social framework of giving and taking reasons (e.g. moral convention or a legal system) the initial motivation of a single human agent could be ignored if his actions and his post-hoc reasoning comply with given social (moral or legal) standards [15]. Reasons within such a post-hoc *"Social Reasoning Model" (SRM)* are not less, but even more important as only they guarantee the coherence of a moral or legal order in an increasingly pluralistic world. The remaining difference is the relative independence of rational reasoning from the motivational impulse to act. Even so, in the long run the inner-subjective or social feedback loop with rational reasons might also change the agents' motivational (emotional) disposition.

This post-hoc SRM is transferable to AI decision processes as *"artificial Social Reasoning Model" (aSRM)*. The black box of an opaque AI system functions like an AI intuition. Following the SRM model, transparency is not needed as long as the system generates post-hoc reasons for its action. Moral and legal accountability and governance could instead be enabled through symbolic or sub-symbolic aSRMs.

A symbolic solution would try to reconstruct (or justify with an alternative argument), the intuitive decision of the black box with deontic logical reasoning applying moral or legal standards. A pluralistic, expressive "normative reasoning infrastructure", such as LogiKEy [3], should e.g. be able to support this process.

A sub-symbolic solution could create an independent (second) neural network to produce reasons for the output of the (first) decision network (e.g. autonomous driving control). Of course, the structure of this "reasoning net" process is again hidden. Yet, if the outcoming reasons coherently comply with prescribed social and ethico-legal standards the lack of transparency in the second black box constitutes less of a problem.

Robust solutions for aSRMs could even seek to tightly integrate and align these two options. Moreover, in both scenarios the introduced feedback loop of giving and taking reasons could be integrated as learning environment (self-supervised learning) for the initial, intuitive layer of autonomous decision making, with the eventual effect that differences at both layers may gradually dissolve.

Allowing various kinds of reasons, SRMs & aSRMs advance normative pluralism and may integrate different (machine-)ethical traditions: deontological, consequentialist and virtue ethics. "Reasonable pluralism" in recent moral and political philosophy defines reasonableness by meta-level procedures like "reflective equilibrium" and "overlapping consensus" [20] or "rational discourse" [10]. Contemporary legal philosophy and theory has enfolded how law could act as democratic real-world implementation of these meta-procedures, structuring public deliberation and argumentation over conflicting reasons [1, 14]. Constructing a pluralist aSRM substantially widens the mostly consequentialist contemporary approaches [4, 8] to machine ethics and moral IAS.

4   Reasonable Machines: Implementation

The implementation of Reasonable Machines requires expertise from different areas: pluralistic normative reasoning, formal ethics and legal theory, expressive ontologies and semantic web taxonomies, human-computer interaction, rule-based systems, automated theorem proving, argumentation technology, neural architectures and machine learning. Acknowledging the complexity of each field, Reasonable Machines research should complement top-down construction of responsible machine architecture with bottom-up developments starting from existing works in different domains. More concretely, we propose a modular and stepwise implementation of our research scheme based on the following modules:

M1: Responsible Machine Architecture. The vision of an aSRM and its parallel to human SRM needs to be further explored to guide and refine the overall architectural design of REASONABLE MACHINES based on respective system components responsible for generating justifications, for conducting compliance checks and for governing the action executions of an IAS.

M2: Ethico-Legal Ontologies. Ethico-legal ontologies constitute a core ingredient to enable the computation, assessment and communication of aSRM-based rational justifications in the envisioned ethico-legal governance components for IASs, and they are also key for black box independent user-explanations in form of rational arguments. We propose the development of expressive ethico-legal upper-level ontologies to guide and connect the encoding of concrete ethico-legal domain-level theories (regulatory codes) [12, 7]. Moreover, we propose the concrete regulatory codes to be complemented with an abstract ethico-legal value ontology, for example, as "discoursive grammar" of justification [16].

M3: Symbolic Reasoning Tools. For the implementation of pluralistic, expressive and paradox-free normative reasoning at the upper-level, the LogiKEy framework [3] can e.g. be adapted and further advanced. This generic approach leverages recent progress in higher-order theorem proving and it shall be complemented and ideally integrated with alternative, more specialized solutions based e.g. on semantic web reasoning, logic programming, answer set programming, and with formalized argumentation for ethical [21] or legal [17] systems design.

M4: Interpretable AI Systems. Sub-symbolic solutions to SRM-based accountability and governance challenge could develop a hidden reasoning net, which might be trained with legal and ethical use-cases. Moreover, techniques in "explainable AI" [9] have to be assessed and, if possible, integrated with the symbolic aSRM tools to be developed in M3 in order to provide guidance to their computations and search processes. The more information can be obtained about the particular information bits that trigger the decisions of the black box systems we want to govern, the easier the corresponding reasoning tasks, i.e. the search for justifications, should become in the associated, symbolic aSRM tool.

M5: Human-Machine Communication & Interaction. The intended aSRM-based justifications generated by the tools developed in M3 and M4 require arguments and rational explanation which are understandable for different AI ecosystems [19], i.e. human users as well as collect decision scenarios between machines and independent verification tools. Here, the development of respective techniques could build on argumentation theory in combination with recent advances towards a computational hermeneutics [6]. An overarching objective of REASONABLE MACHINES is to contribute to trustful and fruitful interaction between human and IASs.

M6: Cloud-based Reasoning Workbench. To facilitate access to the proposed knowledge representation and reasoning solutions, and also to host the ethico-legal theories, a cloud-based reasoning workbench should be implemented. This workbench would (i) integrate the bottom-up construed components and tools from M2-M5 and (ii) implement instances of the top-down governance architecture(s) developed in M1 based on (i). This cloud-based solution could

be developed in combination with, or as an alternative to, more independent solutions based e.g. on agent-based development frameworks [23].

M7: Use Cases and Empirical Studies. The overall system framework needs to be adequately prepared to support changing use cases and empirical studies. Concrete use cases with high ethical and legal potential must be defined and employed to guide the research and development work, as for example the representative issue on self-driving cars [4]. Empirical studies should support and inform the constructive development process. For testing the ethico-legal value ontology in M2, for example, we could try to demonstrate that it can make sense out of the rich MIT Moral Machine experiment data [2]. When its architecture evolves, it would be highly valuable to design a genuine aSRM experiment.

5   Conclusion

The Reasonable Machines vision and research requires the integration of heterogeneous and interdisciplinary expertise to be fruitfully implemented. The cloud-based framework we envision would ideally be widely available and reusable, and it could become part of related, bigger initiatives towards the sharing of critical AI infrastructure (such as the `claire-ai.org` vision towards a CERN for AI). The implementation of the depicted program requires substantial resources and investment in foundational AI research and in practical system development, but it reflects the urgent and timely need for the development of trustworthy AI technology.

The possible outreach of the Reasonable Machines idea is even far beyond an ecosystem of trust. To enable machines to give normative reasons for their decisions and actions means to capacitate them of communicative action [10], or at least to engage in constitutive communication of social systems [18]. The capacity to give and take reasons is a crucial step towards fully autonomous normative (moral and legal) agency. Moreover, our research, in the long run, paves way for interesting further studies and experiments on integrated neuro-symbolic AI architectures and on the emergence of patterns of self-reflection in intelligent autonomous machines.
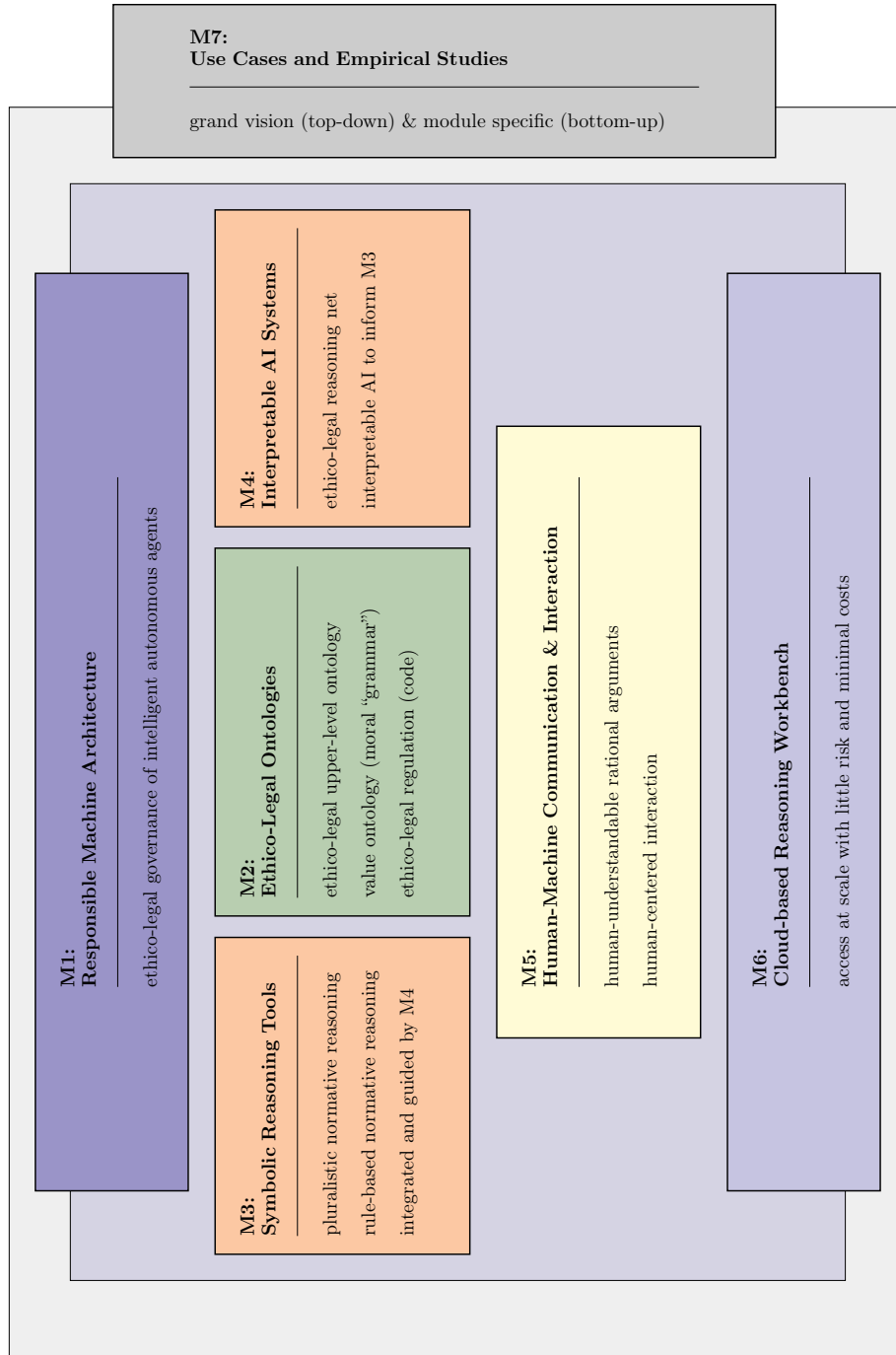
References

1. Alexy, R.: *Theorie der juristischen Argumentation.* Frankfurt/M: Suhrkamp (1978)
2. Awad, E.: The Moral Machine experiment. Nature 563(7729), 59–64 (2018)
3. Benzmüller, C., Parent, X., van der Torre, L.: Designing Normative Theories for Ethical and Legal Reasoning: LogiKEy Framework, Methodology, and Tool Support. Artificial Intelligence (accepted; preprint arXiv:1903.10187) (2020)
4. Bonnefon, J.-F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. Science 352(6293), 1573–1576 (2016)
5. European Commission, *On Artificial Intelligence – A European approach to excellence and trust.* European Commission White Paper, COM(2020) 65 final (2020)

6. Fuenmayor, D., Benzmüller, C.: A Computational-Hermeneutic Approach for Conceptual Explicitation. In: Nepomuceno, A., Magnani, L., Salguero, F., Bares, C., Fontaine, M. (eds.) Model-Based Reasoning in Science and Technology. Inferential Models for Logic, Language, Cognition and Computation, SAPERE, pp. 441–469. Springer, Cham (2019)
7. Fuenmayor, D., Benzmüller, C.: Harnessing Higher-Order (Meta-)Logic to Represent and Reason with Complex Ethical Theories. In: Nayak, A., Sharma, A. (eds.) PRICAI 2019: Trends in Artificial Intelligence, LNAI, vol. 11670, pp. 418–432. Springer, Heidelberg (2019)
8. Greene, J., Rossi, F., Tasioulas, J., Venable, K.B., Williams, B.C.: Embedding Ethical Principles in Collective Decision Support Systems. In: Schuurmans, D., Wellman, M.P. (eds.) Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA, pp. 4147–4151. AAAI Press (2016)
9. Guidotti, R.: A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51(5) (2018)
10. Habermas, J.: *Theorie des kommunikativen Handelns*. Frankfurt/M: Suhrkamp (1981)
11. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol. Rev. 108(4), 814–34 (2001)
12. Hoekstra, R., Breuker, J., Bello, M.D., Boer, A.: LKIF Core: Principled Ontology Development for the Legal Domain. In: Breuker, J., Casanovas, P., Klein, M.C.A., Francesconi, E. (eds.) Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood, Frontiers in Artificial Intelligence and Applications, pp. 21–52. IOS Press (2009)
13. Kahnemann, D.: *Thinking, Fast and Slow*. Farrar, Straus and Giroux (2013)
14. Lomfeld, B.: *Die Gründe des Vertrages: Eine Diskurstheorie der Vertragsrechte*. Tübingen: Mohr Siebeck (2015)
15. Lomfeld, B.: Emotio Iuris. Skizzen zu einer psychologisch aufgeklärten Methodenlehre des Rechts. In: Köhler, Müller-Mall, Schmidt, Schnädelbach, (eds.) Recht Fühlen, pp. 19–32. München:Fink (2017)
16. Lomfeld, B.: Grammatik der Rechtfertigung: Eine kritische Rekonstruktion der Rechts(fort)bildung. Kritische Justiz 52(4) (2019)
17. Lomfeld, B.: Wild Values: Formalized Legal Reasoning With Value Preferences. In: submitted to DEON 2020, (2020)
18. Luhmann, N.: *Soziale Systeme: Grundlage einer allgemeinen Theorie*. Frankfurt/M: Suhrkamp (1984)
19. Rahwan, I.: Machine behaviour. Nature 568(7753), 477–486 (2019)
20. Rawls, J.: *Justice as Fairness: A Restatement*. Cambridge/MA: Harvard University Press (2001)
21. Verheij, B.: Formalizing value-guided argumentation for ethical systems design. Artif. Intell. Law 24(4), 387–407 (2016)
22. Wallach, W., Allen, C.: *Moral machines: Teaching robots right from wrong*. Oxford University Press (2008)
23. Wisniewski, M., Steen, A., Benzmüller, C.: LeoPARD - A Generic Platform for the Implementation of Higher-Order Reasoners. In: Kerber, M., Carette, J., Kaliszyk, C., Rabe, F., Sorge, V. (eds.) Intelligent Computer Mathematics - International Conference, CICM 2015, Washington, DC, USA, July 13-17, 2015, Proceedings, LNCS, vol. 9150, pp. 325–330. Springer, Heidelberg (2015)

M7:
Use Cases and Empirical Studies

grand vision (top-down) & module specific (bottom-up)

M1:
Responsible Machine Architecture

ethico-legal governance of intelligent autonomous agents

M4:
Interpretable AI Systems

ethico-legal reasoning net

interpretable AI to inform M3

M2:
Ethico-Legal Ontologies

ethico-legal upper-level ontology

value ontology (moral "grammar")

ethico-legal regulation (code)

M3:
Symbolic Reasoning Tools

pluralistic normative reasoning

rule-based normative reasoning

integrated and guided by M4

M5:
Human-Machine Communication & Interaction

human-understandable rational arguments

human-centered interaction

M6:
Cloud-based Reasoning Workbench

access at scale with little risk and minimal costs

Fig. 1. Modular structure of REASONABLE MACHINES research.