From Context Mediation to Declarative Values and Explainability

Grzegorz J. Nalepa[1], **Martijn van Otterlo**[2], **Szymon Bobek**[1], **Martin Atzmueller**[2]
[1] AGH University of Science and Technology
{gjn,szymon.bobek}@agh.edu.pl
[2] Department of Cognitive Science and Artificial Intelligence
Tilburg University, The Netherlands
{m.vanotterlo,m.atzmuller}@uvt.nl

Abstract

We argue that there are fruitful (inter)relations between value alignment of AI systems, trust in humans-AI interaction, and declarative representations that make AI systems transparent and explainable. We illustrate matters on two examples and end with directions for research.

1 Introduction

The recent rapid progress of AI enabled "intelligent" technologies to produce new, important and impressive applications. However, AI systems are increasingly covering new areas of decision making that were mostly reserved for humans so far. This raises a number of concerns and questions, i. e., what is a safe degree of autonomy of such systems, who would be *responsible* for their possible failure, or to what degree they can be *trusted*. In fact, there are even more serious, underlying, not so obvious issues, like to what degree do we actually *understand* how some of these systems work, or should they fail, can we explain why they failed, and how we can improve them once they fail. Another question could be whether the design procedure was maybe somehow flawed from the start? An even deeper question would be, which *values* guided the creators of these systems, and which values are actually *embedded* into the AI's code? Ultimately, for a truly general *adaptive* AI system, a crucial question is which values they have *obtained* over their lifetime. While such questions touch upon moral issues, they do relate to *engineering* as well since we *do* need to carefully *design* AI systems that make decisions which have (moral) consequences for people involved. Finally, an even more axiological question is about what should be the *underlying values* for the *creation* of AI systems which are safe, trustworthy, and possibly if not morally good, at least conforming to certain ethical norms.

Clearly, these issues are a current practical concern for engineering AI systems. Legal regulation is also slowly catching up with the rapid progress of AI. Prominent examples are the recent EU regulations on the explainability of AI systems, i. e., the new General Data Protection Regulation, and the "right to explanation", c. f., [Goodman and Flaxman, 2016]. Furthermore, several communities involved with AI systems have started to openly discuss these questions and started to identify *challenges* and *design principles*. The engineering community with the IEEE *Ethically Aligned Design* initiative [1], the robotics community with the EPSRC design principles for robotics systems [2], and the AI community at large with their own *Asilomar principles* [3] which explicitly talk about *judicial* and *failure* transparency: an AI system should be able to *explain* its decisions. All these efforts and concerns are rising regarding both the prospects [4] and benefits where the same concerns are raised about *malicious* [5] *use* of AI.

In our opinion, one of the most important of the underlying problems is the apparent lack of *transparency* of the decision making process in AI systems. From it stem several of the challenges outlined above, specifically regarding *understandability*, *explainability*, and the possibility of incremental improvement (after failures). Furthermore, we believe that this issue of transparency is mandatory to even start addressing the last group of questions, regarding the ethical dimensions of design and operation of intelligent systems.

With this background, in this short paper we assert that transparency in both design and operation of AI systems is needed for them to be explainable, safe, and trustworthy. If we want AI systems to gain a certain level of trust of humans, they need to operate according to human morals and values to some extent. Of course, this does not mean they need to obey some universal ethical system (e.g. Asimov's law's of robotics), but it does mean that it should be possible to demonstrate an alignment of their design and operation to certain ethical codes or sets of norms. Furthermore, we want to stress the role of *humans* in the design and improvement of operation of intelligent systems: *humans are responsible* for the design of AI systems. Moreover, if these systems are designed using machine learning techniques, people are responsible for *teaching* these systems. As we want AI systems to be adaptable and personalized to the needs and expectations of human operators and users, the teaching process does not end in a separate design phase. On the contrary, it continues during the operation and use of the system, and so does the responsibility of humans. To summarize, we claim not only

[1] https://ethicsinaction.ieee.org/

[2] https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

[3] https://futureoflife.org/ai-principles/

[4] https://ai100.stanford.edu/

[5] https://maliciousaireport.com/

In: Proc. IJCAI/ECAI Workshop on Explainable Artificial Intelligence (XAI-18), Stockholm, Sweden, 2018

that humans have to be put in the loop of the design and use of AI systems. We claim, that humans never leave that loop. It is probably a good time for them to realize that.

In this paper, we illustrate our argument with examples that relate it to the design and operation of specific classes of AI systems and models they use. We believe that in order to achieve explainability we need to combine transparent symbolic models for reasoning with specific subsymbolic ones, such as probabilistic reasoning, and machine learning techniques. As such, AI systems have to be heterogeneous by design. Furthermore, we believe that putting the human user in the decision making loop of the system can contribute not only to the system performance, but can also improve the trust the user has in its operation. We claim that if appropriate conditions are met, users are willing to interactively improve how the system works for them. Finally, we demonstrate how specific transparent reasoning techniques can be used to explicitly model some ethical aspects of operation of the system.

2 Context

Transparency and explainability are vital for AI systems in two distinct ways: First, increasingly intelligent and pervasive systems introduced in our society need to be transparent and be able to explain their decisions. Second, explainability will also enhance trust at the user side, and due tothat improve the human-machine interaction performance. Our context is formed by three distinct research topics, discussed below.

Responsible AI The societal consequencs of AI beyond simple privacy and surveillance have become a hot topic in fields ranging from sociology to law under the name *ethics of algorithms* [van Otterlo, 2018b]. Algorithms basically transform *data* into *decisions*, where evidence can be inconclusive, inscrutable or misguided and this can cause many ethical consequences of actions, relating to *fairness*, *opacity*, *unjustified actions*, and *discrimination*. Overall, algorithms have an impact on *privacy* and can have *transformative effects* on *autonomy* [Mittelstadt *et al.*, 2016]. Transparency and the ability to *explain* AI decision making are core requirements for important aspects such as *trust*, *liability*, *responsibility* and *accountability* of algorithms [Diakopoulos, 2016]. In AI and machine learning itself, topics of interest are to ensure *fairness*, *accuracy*, *confidentially* and *transparency* (FACT). Especially transparency has been addressed by *explanation-aware computing* [Atzmueller and Roth-Berghofer, 2011]. Interestingly, the way humans have dealt with ethical issues among people can provide insights into how to deal with ethical consequences of AI. For example, human *codes of ethics* are based on principles of transparency and the ability to explain to the public what norms and values are in particular professions [van Otterlo, 2018a; van Otterlo, 2018b]. Efforts from researchers inside and outside AI enable to understand AI systems and their impact, thereby regulating the (legal) consequences of AI better, and as a result increase trust that the AI will "do the right thing".

Adaptive Human-Machine Interaction Interaction between humans and machines is studied in *human-robot interaction*, *human-computer interaction*, and *interactive machine learning* [Cuayáhuitl *et al.*, 2015]. Until recently, often such interactions were studied with modified *reinforcement learning* [Wiering and van Otterlo, 2012], i. e., how to let the human actor provide guidance to the robotic learner. For example, humans can provide *rewards* or *provide answers to questions from the robot*, and these constructs can be embedded in formal models of goal-oriented tasks. Transparency of the robots behavior, or the interaction, was not an issue so far, although work in *relational robotics* has worked on learning *declarative, probabilistic skill representations* [Moldovan *et al.*, 2012]. Lately, focus has shifted (with influences from responsible AI) to the *value aligment* problem [Abel *et al.*, 2016; Taylor *et al.*, 2017], which is the challenge to construct (adaptive) systems that behave in such a way that they are *aligned* with human values. Narrow interpretations result in various kinds of *inverse* reinforcement learning, but many broader issues are studied to make machine learning systems "safe", including safe exploration, robust generalization over tasks, and avoiding negative side effects of truly intelligent AI which may alter their own reward function at will.

Computing Explanations Recently, the concept of transparent and explainable models has gained a strong focus and momentum in the machine learning and data mining community. Several methods focus on specific model types, e. g., tree-based models [Tolomei *et al.*, 2017]. Also, methods for associative classification, e. g., class association rules [Atzmueller *et al.*, 2018] can be applied for obtaining explicative, i. e., transparent, interpretable, and explainable models [Atzmueller, 2017]. Then, individual steps of a decision can be traced-back to the model, similar to *reconstructive explanations*, c. f., [Wick and Thompson, 1992] on several explanation dimensions [Atzmueller and Roth-Berghofer, 2011]. While the methods sketched above focus on specific modeling methods, there are several approaches for model agnostic explanation methods, e. g., [Ribeiro *et al.*, 2018]. General directions are given by methods considering counterfactual explanation, e. g., [Mandel, 2007; Wachter *et al.*, 2017]. Furthermore, other general methods consider data perturbation and randomization techniques as well as interaction analysis methods, e. g., [Henelius *et al.*, 2017]. Deep neural network models have become a default learning paradigm for huge amounts of data in computer vision, linguistics and reinforcement learning. Their black-box nature has recently triggered several strands of research e.g. focusing on *distilling* learned knowledge into transparent representations such as trees [Frosst and Hinton, 2017] or computing explanations of policy behavior using object saliency maps [Iyer *et al.*, 2018].

Logical, declarative representations, by their nature, provide more options for interpretation [Srinivasan, 2001] and explanation-based reasoning [Atzmueller and Seipel, 2008; van Otterlo, 2009]. With *probabilistic programming languages* these can also be used effectively for machine learning [De Raedt, 2008], e. g., *relational* reinforcement learning [van Otterlo, 2012]. *Comprehensible* machine learning systems can be very useful for transparent and explainable AI.

In the next two sections we will briefly relate to our recent work addressing the selected problems emphasized in the introduction. The first case concerns systems that combine symbolic reasoning with uncertainty handling mechanism supported by the user in an interactive manner.

3 Context Mediation with Human in the Loop

Intelligibility in context-aware systems is an ability of the system to being understood by its users [Lim *et al.*, 2009]. In human-centric systems, such as cognitive advisors this feature is of a key importance, as one of their primary goals is to build trust with a user that helps getting deeper insight into personal preferences and habits. This trust can be built in many different ways, one of which is keeping human in the loop of the decision and possibly learning process. In our recent work, this was achieved by providing an implicit mediation mechanism proposed in [Bobek and Nalepa, 2017a].

The key idea behind mediation is to involve the user into the decision making process, in cases where the output of the inference may be uncertain, and thus may reduce user's trust. In Fig. 1, the architecture of the semantic mediation system for implicit user feedback is shown. The system consists of two main parts: the static knowledge component and the dynamic knowledge component. The static knowledge component provides a semantic representation of the user environment. It allows for communication with the user with the use of concepts that are easily understandable for him. It is an input for the dynamic knowledge component which is responsible for question generation, where questions use concepts from the ontology. These questions aim at obtaining additional information from the user, which can be used to resolve ambiguities or to create new knowledge. Furthermore, the dynamic knowledge component allows for an online mediation between the user and the system.

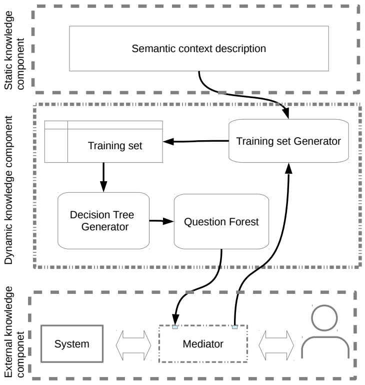
Static knowledge component
Semantic context description
Dynamic knowledge component
Training set
Training set Generator
Decision Tree Generator
Question Forest
External knowledge componet
System
Mediator

Figure 1: Architecture of the implicit semantic mediation system

The system was used by us in a dead-reckoning localization system [Köping *et al.*, 2015], where the implicit mediation was triggered, and where the localization module was highly uncertain about the user position. Based on the ontology that described the environment, it generated easy to answer questions about the most probable user locations. On the one hand this allowed the system to resolve uncertainty, but on the other hand it gave the user better insight on how the decision about his or her location was made. The system we used for localisation was a particle filtering method, a purely mathematical, probabilistic model, and hardly interpretable. We combined it with our implicit mediation system to support the particle filtering algorithm in cases where the estimates of the localisation was very unclear. The localisation of the user was estimated based on the clusters generated from particles. If more than one cluster pointed to different locations on the map, the mediation was triggered.

Uncertainty of knowledge was an inevitable and important challange in the mediation process. It stemed from the fact that sometimes there was no knowledge about the environment (possible location of the user were in rooms which we did not have map of). Furthermore, the user might not be sure about his or her answers. To address this challange, we used a *rule-based* approach combined with *certainty-factors* to allow both uncertainty due to the incompleteness of the model and uncertainty of user answers [Bobek and Nalepa, 2017b].

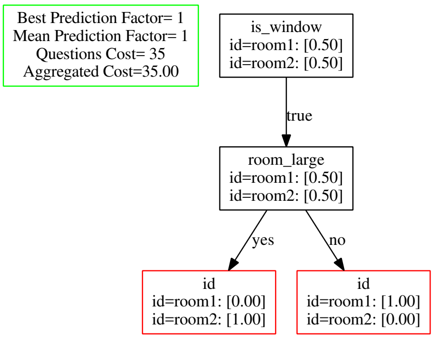

Figure 2: Example of a decision tree from a question forest

The example in Fig. 2 shows the decision tree chosen by the mediation system from the question forest generated for two ambiguous locations. The tree can be then translated into rules of the form presented in Listing 1. Note that the classification accuracy is translated into the form of a certainty factor preceded by the # operator. It allows the user to answer a yes-or-no question with uncertainty as well, and incorporates this uncertainty into reasoning process. Thus, the final result from the reasoning process is also a probabilistic value that can be used by particle filtering algorithm, or with any other subsymbolic method. This extended rule notation is handled by a dedicated inference engine – HeaRTDroid [Bobek and Nalepa, 2017b]. It allows for handling rule uncertainty, but also to operate on attributes related to statistical features of sensor data to deal with machine imprecision.

```
xrule mediate/1: [is_window eq yes,room_large eq yes] ==>
    [id set room2] # 1.0
xrule mediate/1: [is_window eq yes,room_large eq no ]  ==>
     [id set room1] #.1.0
xrule mediate/1: [is_window eq no]==> [id set room1] #-1.0
xrule mediate/1: [is_window eq no]==> [id set room2] #-1.0
```

Listing 1: HMR+ rules representing question tree

We move on to our second case leveraging the combination of declarative models with probabilistic reasoning even more.

4 Declarative Ethical Programs

AI programs should be able to optimize and explain behavior. *Declarative decision-theoretic ethical programs* (DDTEPs) [van Otterlo, 2018a] declaratively specify (and solve) *decision-theoretic problems*, based on the probabilistic programming language (PLL) DT-PROBLOG [Van den Broeck *et al.*, 2010]. Solutions are computed by considering all *possible worlds* modeled by the program. The general idea is to formalize what is known explicitly in the model, and use *reasoning* to compute optimal decisions. DDTEPs fit into logical approaches for ethical (or: value-driven) reasoning [Anderson and Anderson, 2007] but also *relational reinforcement learning* [van Otterlo, 2012] and provides opportunities for *explanation-focused* computations [van Otterlo, 2009] by reasoning over the logical parts of the model.

A (partial) toy example of a DDTEP for a **self-driving car** consists of a decision to either `run_into_wall` (killing the passenger) or a `collision` (killing a pedestrian). We can specify *percepts* for what is in front of the car and a rule that says what happens when a collision is made. The `utility` function defines the *value* of each outcome, making the optimal value -30 (amounting to kill the passenger by steering away).

```
(action)    ?::run_into_wall; ?::collision.
(percepts) in_front_of_car(a).  baby(a).
           in_front_of_car(b). pedestrian(b). ...
(rules)    kill(X) :- in_front_of_car(X), collision.
(values)   utility(run_into_wall, -30).
           utility(kill(X), -20) :- pedestrian(X).
           utility(kill(X), -40) :- baby(X).
```

DDTEPs proved successful for toy ethical domains, e. g., *cake-or-die* and *burning room* [Abel *et al.*, 2016]. There, an agent does not know all values and norms but it can *ask*. This effectively renders the decision problem *partially observable*, requiring *information gathering* first. Through this ”human-in-the-loop” (and value alignment), the AI is told explicitly what the norms are. Specifying *human* values for a DDTEP comes with choices, such as that a `baby` is worth more than a `pedestrian` (or even that they appear on the same *scale*). Such choices can be dependent on social, cultural and other factors as the large-scale experiment *the Moral Machine* [6] aims to investigate, in which people are confronted with ethical *dilemmas* (for autonomous cars) to reveal their intrinsic values.

PLLs typically support *learning*, such that rule parameters (e.g. relative values) can be learned from data instead. Let us take a **fair access in archives** case [van Otterlo, 2018b] where a decision must be made whom to supply with newly disclosed archival material, for example using estimates of a researcher's `authority`. Then, we make the rules *probabilistic*.

```
(rules)  0.1::reach(X):-person(X),social_network(X,small).
         0.9::authority(X):-person(X),h_index(X,high).
     impact(P,T):-topic(T),authority(P).
(action) ?::give(P,T):-person(P),topic(T).
(value)  score(P,T):-give(P,T),impact(P,T).
         utility(score(P,area51),100):-person(P).
```

In this (partial) example, the decision logic dictates that impact depends on authority, which is probabilistically dependent on h-index. The logic is transparent, whereas the numbers can vary with incoming data.

DDTEP open up the black box of algorithms and make decision logic transparent. Still they also allow for machine learning to fill in additional details from data. This general pattern is a solution to value alignment in AI systems in complex domains: i) formalize existing norms and values transparently into a DDTEP, and ii) finetune parts of the program on data. Finetuning is needed because norms and values are never complete and domains are inherently stochastic. Logical formalism such as DDTEP also support *reasoning* about sequential processes, including *explanation-based* reasoning about satisfying norms or obtaining values, opening up the possibility to *ethically explain* its behavior.

5 Outlook and Conclusions

We have argued and illustrated that both declarative representations and the interaction between AI and humans are important factors in getting transparent and explainable systems, also incorporating ideas of *explanation-aware computing* [Atzmueller and Roth-Berghofer, 2011; Atzmueller, 2018]. In the first example it was shown that the combination of an explainable system setup and the opportunity for the user to inspect knowledge and reasoning of the system and to provide feedback, was good for both the trust of the user as well as performance of the system. The second example showed how declarative representations can be used to open up the reasoning of systems that need to make (ethical) decisions and to explain how they get to them. This way again the user can be more engaged in the interaction with the AI and thus more opportunities rise to obtain (optimal) value alignment. Because that is what both examples show: explainability by transparency combined with human users will hopefully result in humans and AI agree more and more on "what is the right thing to do for the AI". Such explainable constructs should be used in the design phase, the interaction (use) phase, but definitely also in the improvement phase of the system over time. As a conclusion, we argue that in order for AI systems to be explainable, they need to be set up as hybrid systems with a good use of declarative knowledge from the start. Explainable systems, we conjecture, will be more trusted by humans and that will increase both performance and value alignment. For achieving these goals, much more research is needed; we want to mention four core directions.

1: Hybrid frameworks that combine declarative knowledge and symbolic reasoning with machine learning. Today AI systems grew far to complex for a single class of models to be sufficient to address problems these systems are supposed to solve. However, there is still an apparent lack of appropriate design approaches for hybrid systems.

2: Integrate explanation-based approaches to reason, provide feedback and support failure analysis. Recent works on explanation generation need to be combined with classic AI approaches to *abductive reasoning*, and *diagnosis*. These mechanisms should not only be used for explanation, but more importantly for provisioning of hypotheses how to ameliorate the system in an understandable way.

3: Use formal analysis techniques to qualify and quantify the performance of the system in terms of value alignment. An important aspect of hybrid frameworks concerns the *formal verification* of their properties on the possibly ethical level, e.g. by expressing all in a (decision-theoretic) logic, it becomes possible to *prove* properties or to analyze *executable ethical specifications* by looking at their potential errors or ethical-logical inconsistencies, or even quantify the ”level”of value alignment achieved so far.

4: Interdisciplinary studies where the performance is analyzed in terms of trust, value alignment, and ethical codes. Clearly AI engineers need to work in integrated yet inter/multidisciplinary teams with knowledgeable researchers from applied ethics, law and sociology to address the issues we outlined. *This is urgent, as AI has already became an important part of our lives, and will continue to be even more so.*

[6] http://moralmachine.mit.edu/

References

[Abel *et al.*, 2016] D. Abel, J. MacGlashan, and M.L. Littman. Reinforcement Learning as a Framework for Ethical Decision Making. In *AAAI Workshop: AI, Ethics, and Society*, 2016.

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

[Atzmueller and Roth-Berghofer, 2011] 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.

[Atzmueller and Seipel, 2008] Martin Atzmueller and Dietmar Seipel. Declarative Specification of Ontological Domain Knowledge for Descriptive Data Mining. In *Proc. International Conference on Applications of Declarative Programming and Knowledge Management (INAP)*. Springer, 2008.

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

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

[Atzmueller, 2018] Martin Atzmueller. Declarative Aspects in Explicative Data Mining for Computational Sensemaking. In *Proc. International Conference on Declarative Programming*. Springer, 2018.

[Bobek and Nalepa, 2017a] S. Bobek and G. J. Nalepa. Uncertain Context Data Management in Dynamic Mobile Environments. *Future Generation Computer Systems*, 66(January):110–124, 2017.

[Bobek and Nalepa, 2017b] S. Bobek and G. J. Nalepa. Uncertainty Handling in Rule-based Mobile Context-Aware Systems. *Pervasive and Mobile Computing*, 39(August):159–179, 2017.

[Cuayáhuitl *et al.*, 2015] H. Cuayáhuitl, N. Dethlefs, L. Frommberger, M. Van Otterlo, and O. Pietquin, editors. *Proceedings of Machine Learning Research*, volume 43. PMLR, 2015.

[De Raedt, 2008] L. De Raedt. *Logical and Relational Learning*. Springer, 2008.

[Diakopoulos, 2016] N. Diakopoulos. Accountability in Algorithmic Decision Making. *CACM*, 59(2):56–62, 2016.

[Frosst and Hinton, 2017] N. Frosst and G. E. Hinton. Distilling a Neural Network Into a Soft Decision Tree. *CoRR*, abs/1711.09784, 2017.

[Goodman and Flaxman, 2016] B. Goodman and S. Flaxman. European Union Regulations On Algorithmic Decision-Making and a "Right to Explanation". *arXiv:1606.08813*, 2016.

[Henelius *et al.*, 2017] Andreas Henelius, Kai Puolamäki, and Antti Ukkonen. Interpreting Classifiers through Attribute Interactions in Datasets. *Proc. 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*, 2017.

[Iyer *et al.*, 2018] R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar, and K. Sycara. Transparency and Explanation in Deep Reinforcement Learning Neural Networks. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 2018.

[Köping *et al.*, 2015] L. Köping, M. Grzegorzek, F. Deinzer, S. Bobek, M. Ślażyński, and G. J. Nalepa. Improving indoor localization by user feedback. In *Proc. International Conference on Information Fusion*, pages 1053–1060, July 2015.

[Lim *et al.*, 2009] B. Y. Lim, A. K. Dey, and D. Avrahami. Why and Why Not Explanations Improve the Intelligibility of Context-aware Intelligent Systems. In *Proc. SIGCHI*, CHI '09, pages 2119–2128, New York, NY, USA, 2009. ACM.

[Mandel, 2007] 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.

[Mittelstadt *et al.*, 2016] 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.

[Moldovan *et al.*, 2012] B. Moldovan, P. Moreno, M. van Otterlo, J. Santos-Victor, and L. De Raedt. Learning Relational Affordance Models for Robots in Multi-Object Manipulation Tasks. In *Proc. ICRA*, 2012.

[Ribeiro *et al.*, 2018] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-Agnostic Explanations. AAAI, 2018.

[Srinivasan, 2001] A. Srinivasan. Four Suggestions and a Rule Concerning the Application of ILP. In S. Dzeroski and N. Lavrac, editors, *Relational Data Mining*, chapter 15, pages 365–374. Springer, 2001.

[Taylor *et al.*, 2017] 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/.

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

[Van den Broeck *et al.*, 2010] G. Van den Broeck, I. Thon, M. Van Otterlo, and L. De Raedt. DTProbLog: A Decision-Theoretic Probabilistic Prolog. In *Proceedings of AAAI*, 2010.

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

[van Otterlo, 2012] 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.

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

[van Otterlo, 2018b] 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.

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

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

[Wiering and van Otterlo, 2012] M.A. Wiering and M. van Otterlo, editors. *Reinforcement Learning: State-of-the-art*. Springer, 2012.