Digital Society (2022) 1:28
https://doi.org/10.1007/s44206-022-00026-y

ORIGINAL PAPER

Learning Domain Ethical Principles from Interactions with Users

Abeer Dyoub[1] · **Stefania Costantini**[1] · **Francesca Alessandra Lisi**[2]

Received: 26 October 2021 / Accepted: 28 October 2022 / Published online: 28 November 2022

Abstract
Drawing from practical philosophy, we argue that AI-based systems could develop ethical decision-making and judgment capabilities by learning from experience. This has inspired our work which combines answer set programming and inductive logic programming to learn domain ethical principles from the interactions with users in the context of a dialogue system.

Keywords AI ethics · Answer set programming · Inductive logic programming · Multi-agent system · Dialogue systems

1 Introduction

Motivation and Background A variety of tools havebeen designed to simplify the interaction between humans and computers. Regardless of their specific form (app, chatbots, etc.), they are in essence dialogue systems, devised for various practical purposes concerning human-machine interaction. Such systems are more and more often based upon artificial intelligence (AI). On the one hand, delegating tasks to these AI-based systems can bring both societal and economic benefits. For example, it may lower costs, increase consistency, and enable new innovative solutions (Taddeo & Floridi, 2018). On the other hand, their use is coupled with ethical challenges. These systems are entering our everyday lives by influencing what we buy, whom we hire, who our friends are, what newsfeed we receive, and even how our children and elderly

✉ Abeer Dyoub
abeer.dyoub@univaq.it

Stefania Costantini
stefania.costantini@univaq.it

Francesca Alessandra Lisi
francesca.lisi@uniba.it

[1] DISIM, University of L'Aquila, L'Aquila, Italy

[2] Department of Computer Science, University of Bari Aldo Moro, Bari, Italy

are cared for. Thus, they must be expected to follow the social and ethical norms of the community in which they are deployed.

Moral decision-making and judgment is a complicated process involving many aspects: it is considered as a mixture of reasoning and emotions. In addition, moral decision-making is highly flexible, contextual, and culturally diverse (Dennis & Clancy, 2022). Since the beginning of this century, there have been several attempts at implementing ethical decision-making into AI agents using different approaches. However, to the best of our knowledge, no fully descriptive and widely accepted model of moral judgment and decision-making exists yet.

Adopting an ethical approach to AI has been attracting a lot of attention in the recent years. The global concern about the ethical behavior of AI-based systems has manifested in a wave of AI-ethics guidelines published in order to maintain social control over technology, with the aim of regulating such systems so that they remain human-centered. Examples of such regulations are EU guidelines for trustworthy AI (Pekka et al., 2018), IEEE ethics (Chatila & Havens 2019), OECD's Recommendation of the Council of Artificial Intelligence (Yeung, 2020), and many others. All these regulations are similar in their key concerns: transparency (sometimes coupled with explainability); justice and fairness; responsibility and accountability; privacy; promote good; autonomy; and related to that (and to accountability) human oversight (Hagendorff, 2020). These regulations did not rely on the terminology of ethics and moral theories, but they use rather general terms, and they are quite concise. Presumably, the tendency of the AI guidelines of being abstract is because of the remoteness of the regulators from the actual AI-based systems. If we try to correlate modern applied ethics and current AI-ethics guidelines, we get the impression that regulators will never have enough time to catch up with technology. With the complexity of AI technologies that are changing rapidly, the strategy seems to be to induce self-regulation by providing abstract guidelines that are to be interpreted w.r.t. the problem at hand, and also to mandate the involvement of the ethicists in the development process (Hagendorff, 2020). We may notice that one cannot analyze or enumerate at the design phase all the unwanted outcomes an AI-based system might produce at run-time. This is due to the open-ended nature of AI. Thus, regulators want AI to avoid bad consequences which designers cannot predict in advance.

The most genuine AI-specific ethical concerns that no other technology exhibit are transparency and human oversight. The point of the issue of human oversight is that AI is transferring control from humans to machines. Since maintaining control requires a hard intellectual work, we would often like to delegate as much control as possible to AI but, at the same time, we want to keep some control over AI to avoid negative outcomes and maintain and preserve the capacity to intervene. AI-ethics regulations cannot prescribe the exact details of a wanted and unwanted AI-based system. Regulators can indeed advise only abstract principles and general recommendations. Thus, developers should be self-regulated. Furthermore, the AI-based systems they produce should be ethical, in the sense that they should be ethically constrained and able to reason about ethics (Héder, 2020). Transparency of the AI-based systems is very important for conducting ethics-based auditing of these systems (Mökander & Floridi, 2021)

Practical philosophy is distinguished from theoretical philosophy in that the latter's aim is contemplation and the understanding of the highest things, while practical philosophy's aim is good action, that is, acting in a way that constitutes or contributes to the good life.

Aristotle believed that ethical knowledge is not only a theoretical knowledge: making the right decision requires in fact more than the knowledge of ethical theory, and in particular requires a sensitivity to the salient features of the situation at hand. In his natural practical philosophy, Aristotle combines logic with observation to make general, causal claims (Hughes, 2004). In Aristotle's practical philosophy (Alesse, 2018), Aristotle aims to formulate a prescriptive ethics, where prescription takes place at the level between universal but highly abstract norms and concrete decision-making; while universal norms tell us in highly abstract terms what to do, prescription answers questions about how to do it. Prescriptive reasoning is a special kind of reasoning which indicates the best thing to do (the most feasible, or the most honorable, depending on circumstances). Aristotle recognized the need to encode practical rules which, although relative to instable and casual reality, can nevertheless be sufficiently constant over time.

Defining a rule means indicating a course of action to solve a practical problem. This course of action has to be both sufficiently specific to meet situational difficulties and sufficiently general and constant over time to offer a code of behavior to be used in similar situations. Furthermore, when we establish rules and prescribe them, we demonstrate the ability of directing not only our own life but also, more importantly, other people's lives. According to Aristotle, it is the experience which provides the principles of each science: logic can only be employed at a later stage to demonstrate conclusions from these starting points.

Hume, generally regarded as one of the most important philosophers that wrote in English, is best known today for his highly influential system of philosophical empiricism, aimed to apply the scientific method to the study of moral philosophy. According to Hume, the only way to improve philosophy is to make the investigation of human nature central and empirical. The problem with ancient philosophy, according to him, was its reliance on "hypotheses," claims based on speculation and invention rather than experience and observation. Hume proposed an empiricist alternative to traditional "a priori" metaphysics. Newton's scientific method provides Hume with a template for introducing the experimental method into his investigation of the mind. Following Newton's example, he argues that we should "reject every system … however subtle or ingenious, which is not founded on fact and observation," and accept only arguments derived from experience. For Hume, all our knowledge comes from experience. Two things are causally connected if the connection could be observed (Botros, 2006).

In Kant's view, the most basic aim of moral philosophy, and so also of his "Groundwork," is practical philosophy that may lead to the creation of moral norms and rules capable of practical implementation. At the heart of Kant's moral philosophy is a conception of reason whose reach in practical affairs goes well beyond that of a Human "slave" to the passions. The fundamental principle of morality in nothing else than the law of an autonomous will which is intrinsic

to the rational agent, it relates to the capacity to freely act according to principles provided by reason (Korsgaard, 2012).

Contribution There is a lack of widely acceptable ethical theories for guiding ethical decision-making. Many proposals in this field developed ethical theories by combining observations from previous literature, commonsense, and experience. However, the tie to actual data has been always weak and doubtful.

Drawing from practical philosophy of Aristotle, Kant, Hume, and other scholars, we argue that AI-based systems could develop ethical decision-making and judgment capabilities by learning from circumstances. In particular, we suggest that useful ethical theories can be induced from real-life cases (examples) in different domains under the supervision of domain ethics' experts. The resulting theories will be novel, and break the gap between the abstract ethical rules and the real life, providing practical guidance for ethical decision-making.

As mentioned above, AI-based systems should respect the AI-ethics regulations and follow the ethical norms (or codes of ethics and conduct) of the community in which they will be deployed. Enforcing such ethical norms in AI-based systems is not an easy task. These norms in fact are mostly abstract and based upon general principles such as confidentiality, accountability, honesty, inclusiveness, empathy, and fidelity that are quite difficult to put into practice in their abstract form. Moreover, abstract principles such as these may contain terms whose meaning may change according to the context. It is difficult to use deductive logic only to address such a problem: it is in fact hardly possible for experts to define fine-grained detailed rules to cover all possible situations. Codes of ethics in their abstract form are very difficult to apply in real situations (Jonsen & Toulmin, 1988). In addition, there are many situations in which obligations might conflict. Learning is needed, in the sense that we need to teach our AI-based systems the codes of ethics and conduct of the domain in which they need to be deployed. Artificial agents might, similar to humans, acquire ethical decision-making and judgment capabilities by implicit processes, in particular via inductive learning (Wallach et al., 2008). Furthermore, with the increase of agents' autonomy, there will be more situations that require morally relevant decisions to be made by an artificial agent interacting with a changing unpredictable environment. Many of these decisions cannot be foreseen in details in advance by a designer.

To tackle these issues, our proposed approach for implementing ethics into AI-based systems combines deductive (rule-based) logic programming and inductive (learning) logic programming in one framework for building our ethical agent (see Sect. 2). As a proof of concept, we introduced in previous work a framework for ethical evaluation of dialogue systems based on the proposed approach, and then implemented this framework as a multi-agent system (c.f. Sect. 3).

2 Automating AI Ethics with Logic-Based AI Techniques

Machine learning (ML) is currently used for critical applications in domains such as healthcare and criminal justice. However, the lack of transparency and accountability of these predictive models can have severe consequences (Adadi & Berrada, 2018).

Conversely, logic-based AI techniques have a great potential to model moral machines due to their inherent comprehensibility (see, e.g., Dyoub, Costantini, & Lisi, 2020).

In Dyoub et al. (2019b, c, d), we proposed a purely declarative approach for automating AI ethics. Our proposal makes use of answer set programming (ASP) as the main knowledge representation and reasoning language, and of inductive logic programming (ILP) for learning the missing ASP rules needed for ethical reasoning. Both ASP and ILP are rooted in the tradition of logic programming. ASP (c.f, e.g., Balduccini et al. (2006), Brewka and Eiter (2016), Dyoub et al. (2018), Erdem et al. (2016), Lifschitz (2017), Lifschitz (2019) for an overview of ASP and its applications) is a successful purely declarative non-monotonic reasoning paradigm. ASP has been our choice because ethical rules are by their very nature default rules, which means that they tolerate exceptions. This in fact nominates non-monotonic logics, which simulate common sense reasoning, to be used for formalizing different ethical conceptions. There are the many advantages of ASP including its expressiveness, flexibility, extensibility, ease of maintenance, and readability of its code. In addition, the existence of free inference engines ("solvers") to derive consequences of different ethical principles automatically can help in precise comparison of ethical theories, and makes it easy to validate our models in different situations.

ILP (Law et al., 2019) is a kind of ML aimed at learning logic programs. As opposed to (statistical) ML methods, ILP does not require huge amounts of training examples and produces interpretable results. ILP is known for its explanatory power, and clauses of the generated rules can be used to formulate an explanation for the choice of certain decisions over others. So, ILP appears to be particularly suitable and promising for automating AI ethics, where the scarcity of examples is one of the main challenges, and the comprehensibility of the output is indispensable. Comprehensibility of logic-based representations is in fact one of their most recognized advantages. Thus, the resulting agents are transparent by design.

The proposed approach is based on the elaboration of facts extracted from codes of ethics and conduct, formal documents proper of the given domain or organization, and from real-life situations concerning pertinent ethical decision-making and judgment. These facts are used to elicit rules for ethical reasoning. Thus, the approach is general enough to produce ethical reasoning rules for any domain.

Initially our AI-based agent will have in its knowledge base, the domain knowledge, a small ethical background knowledge limited to the ethical codes and norms that could be encoded deductively (hand-coded) using ASP. The missing ethical rules are learned by our agent incrementally overtime from interaction with users, during training, testing, and operation phases, under the supervision of the domain ethicist. The newly learned (generated) ethical rules are added to the agent's knowledge base to be used for ethical reasoning about future cases. Practical ethical principles are rules of behavior, in other words, rules that help us to decide what is an ethical action, and what is not ethical. In addition, they help us to ethically judge and evaluate the behavior of others. Thus, any ethical system, i.e., any consistent set of ethical principles, needs the definition of an associated decision making procedure. Considering the domain of interest, we want to describe these decision-making procedures in a purely declarative way. Using the ASP formalism, it is possible to

model ethical rules explaining the status of a certain case situation (or a set of similar cases).

During the training phase, the trainer enters a series of cases (examples), along with the ethical evaluation of the examples in each scenario. The system remembers the facts about the narratives provided by the trainer, and learns to form ethical evaluation rules according to the facts which are recorded in the story context and background knowledge. The learnt ethical rules needed will dictate the ethical behavior of our agent. When the agent faces a new case scenario, it will check its knowledge base for an ethical rule to use for reasoning about the current case. If the agent does not have the needed rule for ethical evaluation of the case at hand, it will start the learning process to learn/revise the missing rule with the help of domain ethicist.

In Anderson et al. (2005), the authors used ILP to learn rules to help decide between two or more available actions based on a set of involved ethical "prima facie" duties (in their work used the prima facie duties of biomedical ethics: autonomy, beneficence, nonmalficence). So, their approach can be applied to choose the most ethical action when we have specific clear ethical duties involved and to do so we need to assign weights of importance (priority) to these duties for each available action, then the system computes the weighted sum for each action, and the one with highest weighted sum is the best action to do. In this approach, it is not really clear what is the criterion underlying the assignment of weights to duties (we doubt whether we can really quantify the importance of ethical duties on a grade from 2 to −2 as done in these works). Then, it is not clear whether the generated rules can be refined incrementally over time. Instead, in our approach, we use ILP to generate rules for ethical evaluation of actions based on different facts extracted from cases. In other words, ILP is used to learn the relation between the evaluation of an action to be ethical or unethical and the related facts in the case scenario. To this end, different facts are extracted from the case scenario and our system tries to find the relation between these facts and the conclusion (ethical or unethical or probably unknown). We think that our approach is more general, can be used to generate ethical rules for any domain (and/or elaborate existing ones), and does cope with the changes of ethics over time because of the use of nonmonotonic logic and incremental learning.

3 A Framework for Ethical Evaluation of Dialogue Systems

In the work Dyoub et al. (2019a), we proposed a framework for ethical evaluation of dialogue systems, and then implemented this framework as a multi-agent system (MAS) (Dyoub, Costantini, Letteri, & Lisi, 2021; Dyoub, Costantini, Lisi, & De Gasperis, 2020). The resulting system, called *EthicalEvalMAS*, is a pilot system aiming in the first place to test the previously proposed ethical evaluation approach, and constitutes a step towards building practical ethical machines. The proposed framework acts as a separate ethical layer that can be integrated within any dialogue system. Online customer service was the application domain chosen to conduct the experiments with the proposed system. The system was trained and tested using a very small dataset of 100 examples created manually (invented dialogue scenarios). Each example is composed of a set of facts extracted from the dialogue text and

a label (ethical/unethical). In the context of an online customer service dialogue system, we intended to ensure an ethical behavior from the chatting agent (human/artificial). Online customer service agents are in fact monitored for ethical violations by the proposed architecture. In order to achieve this overall goal, the MAS is composed of a group of agents, where each one is responsible for a specific sub-task in the overall ethical monitoring task. The online customer service environment in this work consists of clients, online customer service agents (human/artificial), and software agents. A client interacts with the system via a chatting point interface, where she/he can write her/his requests (questions), and receive answers. Answers to the client's requests are given by the online customer service agent. Software agents in the environment are client agent (CA), chatting agent (ChA), text extractor agent (TEA), text-ASP translation agent (TATA), ethical evaluation agent (EEA), and monitoring agent (MA). The ethical evaluation agent has two primary goals: (1) to generate an ethical evaluation of the online customer service agent's answers using the ASP reasoning module, which utilizes the current case facts, and the background knowledge (BK) from the knowledge base (KB) to elaborate the evaluation. (2) Learning the ethical rules needed for ethical evaluation, and saving them into its KB, in case the ASP reasoning module is not able to give an evaluation. The MA agent is currently responsible only for alerting the CA agent for ethical violations (the role of this agent can be extended to practice more control over the CA agent). For more details and examples, please refer to (Dyoub, Costantini, Letteri, & Lisi, 2021; Dyoub, Costantini, Lisi, & Letteri, 2020; Dyoub, Costantini, Lisi, & Letteri, 2021).

The implemented MAS model is a pilot project, which is still in its preliminary phases with many limitations, the main one being the unavailability of a big enough dataset for training. In fact, this is one of the major challenges in the ethical domain in general. This is due to two reasons. First, the field of machine ethics is a new field with very little pre-existing research work. Second, the sensitivity of the ethics domain makes it very difficult to acquire data due to privacy reasons. However, we intend in the future to adapt the MAS system that we created for testing, for the creation of training datasets. It is worth mentioning that we developed and published a web application (remained available for 1 year for voluntary participation) meant to collect data, regarding unethical scenarios, for producing a big dataset to train our online customer service chatbot. Unfortunately, from this web application, we obtained only a couple of useful scenarios. Other limitations are related to the natural language translation module; the development of a more effective translator is in our future plans. Furthermore, the system is not yet fully autonomous, human-in-the-loop is still needed. Anyway, we believe that in principle the human-in-the-loop should always be there to avoid any negative consequences and maintain the possibility to intervene if needed. Machines do not possess a "will" of their own nor understand the concept of freedom, and how to attain it by adopting principles that will develop inner and outer autonomy of the will. A machine has no self-determining capacity that can make choices between varying degrees of right and wrong. Simulating the devised framework by means of a MAS helped us to get better insights into the dynamics of a corresponding real-world system, and to assess the practical challenges and limitations of building such a system.

4 Discussion and Conclusions

Implementing ethics or making ethical decision-making computable provides many advantages: (I) adding ethical dimensions to the AI-based systems that are becoming increasingly autonomous, which leads to avoiding possible harmful behavior from them; (II) help us to better understand and advance the study of ethical theories; and (III) develop a decision procedure for ethical theories, which is an essential problem with most of them, especially those that involve multiple conflicting rules.

In an ill-defined domain like the machine ethics domain, it is infeasible to define abstract codes in precise and complete enough terms to be able to use deductive problem solvers and to apply them correctly. A combination of deductive (rule-based) and inductive (case-based learning) is in our opinion in order. Integrating deductive and inductive logic–based reasoning (ASP and ILP) for modeling ethical agents provides many advantages: increases the reasoning capability of our agent, promotes the adoption of hybrid strategy that allow both top-down design and bottom-up learning via context sensitive adaptation of models of ethical behavior, and allows the generation of rules with valuable expressive and explanatory power which equips our agent with the capacity to give an ethical evaluation and explain the reasons behind this evaluation. In summary, our method supports transparency and accountability of such models, which facilitates instilling users' confidence and trust in our agent.

ILP algorithms, unlike neural networks (NN), output rules which are comprehensible by humans and can provide an explanation for predictions on a new data sample. Furthermore, in NN, if prior knowledge (background knowledge) is extended, then the entire model needs to be re-learned. Finally, no distinction is made between exceptions and noisy data in these methods. This makes ILP particularly appropriate for scientific theory formation tasks in which management of noisy data and exceptions and comprehensibility of the generated knowledge is essential.

Providing explanations of a system's decisions is fundamentally linked to its reliability and trustworthiness. An ASP program and its output models contain both the output and the justification for the given output, which can be easily shown to the user. No need for further processing to generate the explanations for the users. The explanations are already part of the output model.

Pro-social rule breaking (PSRB) behavior (Morrison, 2006) is an intentional violation of rules to promote the welfare of one or more stakeholders. Morrison's research found that 60% of rule-breaking cases are pro-socially motivated. Our approach helps to implement PSRB-capable ethical governors (modules) that can learn PSRB behavior on the basis of the experiences of virtuous experts (i.e., in the case of autonomous vehicles, virtuous drivers). Ramanayake and Nallur (2022) suggests PSRB-capable ethical governors to enhance ethical abilities of current AI systems.

The developed MAS acts as a separate ethical component (ethical layer) for ethical evaluation, which provides many advantages from an engineering point of view: (I) The ethical component has access to all data used for ethical evaluation, and uses this data to provide justifications for a given ethical evaluation to humans, which leads to accountability. (II) The possibility to adapt the ethical component to changes in circumstances and needs. In addition, the possibility of implementing more than one

version of the ethical component on the same agent. (III) The possibility to check and verify the functionality of the ethical component independently from the operations of the autonomous agent. (IV) The re-usability and standardization. Having a separate component for ethical evaluation gives us the possibility to standardize this ethical component, which will have the advantage of avoiding the need to re-invent ethical components that fit for a large number of agents' architectures.

In a MAS model, it is very easy to incorporate modifications in the behavior of agents, by adding behavioral rules which operate at the agent level. It is also possible to dynamically add new agents with their own behavioral model, which interact with the already-defined agents, without having to recompile or even re-initiate the system. Extensibility is in fact one of the most powerful features of agent-based systems. The way in which agents are designed makes them also easier to be reused than objects.

The ethical evaluation of the proposed MAS system is based on the facts extracted from the case scenario, and their relation to the codes of ethics and conduct, which results in a set of ethical evaluation rules, against which one can evaluate the behavior of an agent. These rules are used to decide whether the agent action is ethical/unethical. Evaluating the decidability and completeness of the generated rules is an open issue, and is a matter of further experiments and evaluation. The system needs substantial improvements and comprehensive testing before it is ready for market. Furthermore, issues such as scalability and fault-tolerance are paramount to the successful operation of any application, and even more so when the application deals with sensitive issues like ethics.

Potential case studies A potential case study, on which we are currently working, is ethical care robots (we have already obtained some very preliminary results, a detailed discussion is out of the scope of this article). In this work, we show how care robots can learn logical ethical rules of behavior, from experience, under the supervision of a human ethical teacher. The ethical principles adopted in the medical domain are beneficence, non-maleficence, autonomy, and justice (Beauchamp and Childless 1991). They are very abstract principles, subject to interpretations, and very hard to implement in concrete situations. It is also very hard to define practical detailed rules to cover all possible situations that a care robot might encounter. We considered as a case study a care robot working in a nursing house, where it should support elderly persons in their daily life by carrying out some services for them. From the scenarios that this care robot encounters, it learns, with the help of its human ethical teacher, the ethical rules of behavior of the nursing house. Then, it applies these rules to choose the ethical action to perform in a certain situation. With care robots still in a stage of relative infancy, the discovery of new ethical issues is likely to continue. Robots should have the ability to learn continuously from these emerging cases and build their guiding principles and ethical standards. Another case study that we are planning to consider is conversational agents in some healthcare domain and personal healthcare assistants.

Finally, we believe that the proposed framework, and its realization as a MAS model, has a great potential for future design and implementation of ethical machines in different domains. More generally, we are convinced that the ethical

behavior of autonomous agents in different domains should be guided by explicit and transparent evolving ethical rules.

Author Contribution A novel logic-based hybrid approach for ethical decision-making and judgment which integrates reasoning and learning in one framework. In this approach, the ethical agent learns the domain ethical principles from interactions with users and exploits these learned ethical principles to make ethical decisions in the future. A framework, implemented as a multi-agent system, for ethical monitoring and evaluation with application in online customer service domain.

Funding Open access funding provided by Università degli Studi dell'Aquila within the CRUI-CARE Agreement.

Data Availability Data are available from the authors upon reasonable request.

Materials and/or Code Availability NotApplicable (or all needed links were provided in the text).

Declarations

Ethics Approval Not applicable.

Informed Consent Not applicable.

Competing Interests The authors declare no competing interests.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alesse, F. (2018). *Aristotle on prescription: Deliberation and rule-making in Aristotle's practical philosophy*. Brill.

Anderson, M., Anderson, S. L., & Armen, C. (2005). MedEthEx: Toward a medical ethics advisor. In *Caring machines: AI in eldercare, papers from the 2005 AAAI Fall Symposium, Arlington, Virginia, USA, November 4–6, 2005* (vol. FS-05-02, pp. 9–16). AAAI Technical Report. AAAI Press, USA. Retrieved 2018, from https://www.aaai.org/Library/Symposia/Fall/fs05-02.php

Balduccini, M., Gelfond, M., & Nogueira, M. (2006). Answer set based design of knowledge systems. *Annals of Mathematics and Artificial Intelligence, Springer, 47*(1–2), 183–219. https://doi.org/10.1007/s10472-006-9026-1

Beauchamp, T. L., & Childless, J. F. (1991). Principles of biomedical ethics. *International Clinical Psychopharmacology, 6*(2), 129–130. https://doi.org/10.1001/jama.1984.03340360075041

Botros, S. (2006). *Hume, reason and morality: A legacy of contradiction*. Routledge.

Brewka, G., & Eiter, T. (2016). M.T. (eds.) Answer set programming: Special issue. *AI Magazine*, *37*(3).

Chatila, R., & Havens, J. C. (2019). The IEEE global initiative on ethics of autonomous and intelligent systems. In *Robotics and well-being* (pp. 11–16). Springer.

Dennis, M. J., & Clancy, R. F. (2022). Intercultural ethics for digital well-being: Identifying problems and exploring solutions. *Digital Society, 1*(1), 1–16.

Dyoub, A., Costantini, S., & De Gasperis, G. (2018). Answer set programming and agents. *Knowledge Eng. Review, 33*, e19. https://doi.org/10.1017/S0269888918000164

Dyoub, A., Costantini, S., & Lisi, F. A. (2019a). An approach towards ethical chatbots in customer service. In *Proceedings of the 6th Italian Workshop on Artificial Intelligence and Robotics co-located with the XVIII International Conference of the Italian Association for Artificial Intelligence (AI*IA 2019), Rende, Italy, November 22, 2019, CEUR Workshop Proceedings*, (vol. 2594, pp. 1–5). CEUR-WS.org. Retrieved 2019, from http://ceur-ws.org/Vol-2594

Dyoub, A., Costantini, S., & Lisi, F. A. (2019b). Learning answer set programming rules for ethical machines. In *Proceedings of the Thirty Fourth Italian Conference on Computational LogicCILC, June 19–21, 2019, Trieste, Italy*. CEUR-WS.org. Retrieved 2019, from http://ceur-ws.org/Vol-2396/

Dyoub, A., Costantini, S., & Lisi, F. A. (2019c). Towards an ILP application in machine ethics. In *Inductive Logic Programming - 29th International Conference, ILP 2019, Plovdiv, Bulgaria, September 3–5, 2019, Proceedings, Lecture Notes in Computer Science* (vol. 11770, pp. 26–35). Springer, Netherlands. https://doi.org/10.1007/978-3-030-49210-6

Dyoub, A., Costantini, S., & Lisi, F. A. (2019d). Towards ethical machines via logic programming. In *Proceedings 35th International Conference on Logic Programming (Technical Communications), ICLP 2019 Technical Communications, Las Cruces, NM, USA, September 20-25, 2019, EPTCS* (vol. 306, pp. 333–339). https://doi.org/10.4204/EPTCS.306.39

Dyoub, A., Costantini, S., & Lisi, F. A. (2020). Logic programming and machine ethics. In *Proceedings 36th International Conference on Logic Programming (Technical Communications), ICLP Technical Communications 2020, (Technical Communications) UNICAL, Rende (CS), Italy, 18-24th September 2020, EPTCS* (vol. 325, pp. 6–17). Retrieved 2020, from https://doi.org/10.4204/EPTCS.325.6

Dyoub, A., Costantini, S., Lisi, F. A., & De Gasperis, G. (2020). Demo paper: Monitoring and evaluation of ethical behavior in dialog systems. In *Advances in practical applications of agents, multi-agent systems, and trustworthiness. The PAAMS Collection - 18th International Conference, PAAMS 2020, L'Aquila, Italy, October 7-9, 2020, Proceedings, Lecture Notes in Computer Science* (vol. 12092, pp. 403–407). Springer, Netherlands. https://doi.org/10.1007/978-3-030-49778-1_35

Dyoub, A., Costantini, S., Lisi, F. A., & Letteri, I. (2020). Logic-based machine learning for transparent ethical agents. In F. Calimeri, S. Perri, E. Zumpano (Eds.), *Proceedings of the 35th Italian Conference on Computational Logic - CILC 2020, Rende, Italy, October 13-15, 2020, CEUR Workshop Proceedings* (vol. 2710, pp. 169–183). CEUR-WS.org. Retrieved 2020, from http://ceur-ws.org/Vol-2710/paper11.pdf

Dyoub, A., Costantini, S., Letteri, I., & Lisi, F. A. (2021). A logic-based multi-agent system for ethical monitoring and evaluation of dialogues. In A. Formisano, Y. A. Liu, B. Bogaerts, A. Brik, V. Dahl, C. Dodaro, P. Fodor, G. L. Pozzato, J. Vennekens, & N. Zhou (Eds.), *Proceedings 37th International Conference on Logic Programming (Technical Communications), ICLP Technical Communications 2021, Porto (virtual event), 20–27th September 2021, EPTCS* (vol. 345, pp. 182–188). https://doi.org/10.4204/EPTCS.345.32

Dyoub, A., Costantini, S., Lisi, F. A., & Letteri, I. (2021). Ethical monitoring and evaluation of dialogues with a MAS. In: S. Monica, F. Bergenti (Eds.), *Proceedings of the 36th Italian Conference on Computational Logic, Parma, Italy, September 7–9, 2021, CEUR Workshop Proceedings* (vol. 3002, pp. 158–172). CEUR-WS.org. Retrieved 2021, from http://ceur-ws.org/Vol-3002/paper13.pdf

Erdem, E., Gelfond, M., & Leone, N. (2016). Applications of answer set programming. *AI Magazine, 37*(3), 53–68.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120.

Héder, M. (2020). A criticism of AI ethics guidelines. *Információs Társadalom: Társadalomtudományi Folyóirat*, *20*(4), 57–73. https://doi.org/10.22503/inftars.XX.2020.4.5

Hughes, G. J. (2004). Aristotle on ethics. *Tijdschrift Voor Filosofie, 66*(1), 176–176.

Jonsen, A. R., & Toulmin, S. E. (1988). *The abuse of casuistry: A history of moral reasoning*. Berkeley: Univ of California Press, USA.

Korsgaard, C. M. (2012). *Kant: Groundwork of the metaphysics of morals*. Cambridge University Press.

Law, M., Russo, A., & Broda, K. (2019). Logic-based learning of answer set programs. In: M. Krötzsch, D. Stepanova (Eds.), Reasoning Web. Explainable Artificial Intelligence - 15th International Summer School 2019, Bolzano, Italy, September 20–24, 2019, Tutorial Lectures, *Lecture Notes in Computer Science* (vol. 11810, pp. 196–231). Springer. https://doi.org/10.1007/978-3-030-31423-1_6

Lifschitz, V. (2017). Achievements in answer set programming. *Theory and Practice of Logic Programming*, *17*(5-6), 961–973. https://doi.org/10.1017/S1471068417000345

Lifschitz, V. (2019). *Answer set programming*. Springer. https://doi.org/10.1007/978-3-030-24658-7

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds Mach., 31*(2), 323–327. https://doi.org/10.1007/s11023-021-09557-8

Morrison, E. W. (2006). Doing the job well: An investigation of pro-social rule breaking. *Journal of Management, 32*(1), 5–28.

Pekka, A., Bauer, W., Bergmann, U., Bieliková, M., Bonefeld-Dahl, C., Bonnet, Y., Bouarfa, L., et al. (2018). The European commission's high-level expert group on artificial intelligence: Ethics guidelines for trustworthy AI. *Working Document for stakeholders' consultation* (pp. 1–37). Brussels.

Ramanayake, R., & Nallur, V. (2022). Pro-social rule breaking as a benchmark of ethical intelligence in socio-technical systems. *Digital Society, 1*(1), 1–6.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752.

Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI Soc., 22*(4), 565–582. https://doi.org/10.1007/s00146-007-0099-0.

Yeung, K. (2020). Recommendation of the council on artificial intelligence (OECD). *International Legal Materials, 59*(1), 27–34.