# New-Generation AIs Reasoning about Norms and Values

## Subproject Description for the Logic for New-Generation AI project

Réka Markovich[a] Amro Najjar[a] Leendert van der Torre[a,b]

[a] *University of Luxembourg*

[b] *Zhejiang University*

**Abstract**

The recent rapid evolution of artificial intelligence and its widespread application in a multitude of domains led to the emergence of new heterogeneous systems where humans cohabit with software agents. The logics needed in these new-generation systems with intelligent behavior (AIs) have to accommodate reasoning about norms and values: these considerations drive humans' everyday life and decisions, we expect artificial intelligence tools to operate in our society taking these considerations very much into account. Some of these norms, like the legal ones, are mostly explicit, some, however, just like the values behind them, are not: the moral, social and cultural norms, while crucially affecting what people do and why, are usually not written, especially not precisely phrased, sometimes even not consciously reflected. In order to make the logics used for normative reasoning in artificial intelligence legitimate, next to the mathematical and technological requirements of such a formal system, we have to gather insights about these tacit or vague norms and values people reason with. In order to do so, next to the usual theoretical methodologies of developing new formal systems, data-driven and experimental methodologies with people, agents and texts are also needed revealing both the norms and values, and the ways people reason with them in different cultures. Engaging in this wide range of methodologies realizing a cross-disciplinary endeavor serves the purpose of gaining a comprehensive overview of what reasoning with norms and values is suitable for the new generation of AIs.

*Keywords:* normative reasoning, deontic logic, reasoning with values, data-driven approach, cross-cultural approach, cross-disciplinarity

## 1 Introduction

'Artificial intelligence' is used to describe the field studying and engineering intelligent agents [36], and nowadays it used as well as a synonym for intelligent agents themselves, either software agents or physical robots. In the new generation AIs, computer vision and machine learning play prominent roles, while at the same time being capable to represent knowledge and reason about it. Moreover, these new generation AIs are characterized by natural interaction with humans, and having significant social impact in the real world.

There are some key requirements that have been identified for these new generation AIs. One is that the natural interaction with humans implies that they can explain their behavior in an intelligent way to the humans they interact with, as well as to other AIs [4,22]. Another requirement, also implied by operating within our society interacting with humans, is that their actions be within the same normative framework as that of the humans'. Legal, moral and cultural norms and values provide the framework for our society defining the normative space of our actions, we expect the AIs' actions also to be within it, and their ability to reason with these norms and values to be a constituent of their intelligence [26]. In order to make sure that the decisions made by AIs are equipped with normative considerations, it is necessary to obtain and represent the existing legal and ethical norms of our societies. Achieving this goal requires facing and overcoming several challenges, though. While legal norms are mostly written, moral ones are almost never, the values and preferences over them are often not even consciously reflected. Thus, a methodology should exist for extracting norms, values and preferences from texts and from people's judgements and behavior.

Next to the norms themselves, preferences over these norms and values construct the basis for normative reasoning. The task of reasoning about values requires a parallel and highly connected endeavor when thinking about norms. Value and preference are everywhere in our daily life, and the notions of value and preference play key roles in research fields such as moral philosophy, and psychology.

A resilient reasoning system should be non-monotonic to handle conflicts between norms and values and should be able to reason on the meta-level too where more than one normative systems could be applied [25]. Also, such a reasoning system's accuracy should be verifiable. The technologies of cognition-driven and data-driven to establish a new model of value and normative reasoning research, and to combine value and ethical principle mining technology with the data-driven model in the field of legal reasoning. Moreover, we can model how an expected output, such as a set of ethical principles or a value preference, can be enforced given a multi-agent normative system and to investigate the computation complexity of the model.

The norms and values are different in the different societies [7], within those, they differ too in the different communities, and different stakeholders of a given AI tool might have different normative considerations even within the same community. Investigating and taking these differences into account while developing reasoning systems leads to a comparative study as an output of such a research project. This study is on similarities and differences of norms and values in the context of multi-culture and compare the ethical principles and normative systems between China and Europe, resulting from collecting data and cases of artificial intelligence related to norms and ethics in China and Europe, and evaluating and improving the algorithms of the models with instantiated data.

In the LNGAI project we adopt a non-monotonic logic to establish a value and normative reasoning model. We establish a data-driven ethical decision-making model. We also establish a multi-agent system based on norms and values and verify the accuracy of the system. Finally, in the LNGAI project we make a comparative study on similarities and differences of norms and values in artificial intelligence between China and European countries. This paper's layout is aligned to this research plan. During the discussion, we start from the following rough definitions:
Def 1. AI(s) = system(s) with intelligent behavior.
Def 2. New generation AI(s) = AI(s) (1) based on perception, representation and reasoning, and learning, (2) displaying natural interaction, and (3) having social impact.
Def 3. Explainable AI(s) = AI(s) that can explain behavior in an intelligent way to humans and other AIs.
Def 4. Legal, ethical, moral, social, cultural norms, normative system for AIs = individual and collective expressions of what is usual, typical or standard from the perspective of some discipline, institutions, organization, society or culture.
Def 5. Value, preference for AIs = individual and collective judgments of what is important in life.
Def 6. Deontic logic, preference logic for AIs = formal languages that can be used for the logical analysis of normative AIs, describing logical relations between the AIs and their rights and duties

## 2 Logic for Values and Norms

The prominent group of formalisms in normative reasoning is deontic logic. The foundations of deontic logic was created in the 1950's [43], embedded in the modal logic tradition, and its later variant is still referred to as standard deontic logic [15]. In this tradition, the main emphasis is on the obligations and permissions and the basis of the semantics are possible worlds. These systems, while being rather intuitive, suffer from paradoxes, and fundamentally monotonic. In the last decades, another tradition emerged in normative reasoning, closer to a rule-based approach, which explicitly refers to the norms themselves [3,29,33]. Agent-based modelling, for instance, BDI (belief-desire-intention) [34] can also be extended with the notion of obligation resulting in BOID models [12].

In a dynamic and open environment, the normative system and the agent's value system are unknown and changing. Also, normative systems often operate with exceptions and has to handle possible conflicts within the system. Not surprisingly, non-monotonic logics emerging in the 1980's in the filed of computer science [37] was subsequently applied to normative reasoning in the form of logic programming [5,39] and default logic [35,18]. But these formal languages based on computer programming are lacking in expressing ability and portraying diverse environments. In recent years, the formalism of argumentation theory and causal reasoning in norm and value reasoning has gradually attracted attention [9,24]. However, this work is still relatively preliminary. Thus, how to use the latest theories and methods of non-monotonic logic, formal argumentation systems and causal reasoning to carry out norms and values in an open and dynamic environment for reasoning with norms and values has not yet been systematically studied.

The so-called LogiKEy: Logic and Knowledge Engineering Framework and Methodology is also a recent development for designing and engineering ethical/legal reasoners, normative theories and deontic logics [10]. LogiKEy provides an integrative framework and methodology for using and developing new logics for reasoning with norms making it possible to continuously check the results against the ethical or legal theory we start from. Thus, LogiKEy is an essential asset when thinking about how the new generation AIs should reason about norms. It was created from an application-aiming approach, and it has the potential to revolutionize the area of deontic logic by addressing directly the decades-long challenge of how deontic logics and normative theories can be used in computer science applications. In the meantime, though, LogiKEy doesn't shift the focus from the theoretical basics: its pivotal property is the overarching nature. The unifying formal framework LogiKEy offers is based on semantical embeddings of deontic logics, logic combinations and ethical or legal domain theories in expressive classical higher-order logic (HOL). This meta-logical approach enables the provision of powerful tool support in LogiKEy: off-the-shelf theorem provers and model finders for HOL are assisting the LogiKEy designer of ethical intelligent agents to flexibly experiment with underlying logics and their combinations, with ethical or legal domain theories, and with concrete examples at the same time.

The task of reasoning about values requires a paralel and highly connected endeavor when thinking about norms. Value and preference are everywhere in our daily life, and the notions of value and preference play key roles in research fields such as moral philosophy, psychology, decision theory, game theory, and social choice. Value is often expressed as monadic preference, such as "I prefer to go to the West Lake", "the committee prefers to make its decisions available on the website", "it is preferred to be honest", or "people prefer symmetric faces". Preference is usually expressed in terms of a comparison between two objects or situations, using comparative statements such as "If we are served duck, then I prefer rice over noodles", or "peace is preferred over war".

Preference logic formalises reasoning about value and preference statements. Reasoning about preference is challenging, both conceptually and computationally. Some of the conceptual challenges are the aggregation of preferences, the change of preference, or the definition of the ceteris paribus proviso. Examples of computational challenges are efficient querying of preference, preference elicitation, communication of preference, and non-monotonic reasoning about preference.

Traditionally the emphasis in reasoning about value and preference was on intrinsic preference, but more recently the emphasis has shifted to extrinsic preference. Roughly, extrinsic preferences are based on reasons, while intrinsic preferences are not. For example, Von Wright's logic of preference [46] is explicitly restricted to intrinsic preferences in the sense that he considers preferences not having an extrinsic reason or motivation, whereas in deontic logic the preference ordering is derived from a set of norms in norm-based semantics [28]. In general, the logic of extrinsic preference makes the reasons explicit, and thus reasons about both preference and the reasons for preference. In moral reasoning, value may be one of the reasons behind norms.

As a consequence of making the reasons for extrinsic preference explicit, the formalisation of value and preference change has become a central concern for preference logic as well. For example, extrinsic preferences may be changed by commanding or promulgating norms. In general, extrinsic preferences change when the reasons change, or when the priorities amongst these reasons change, whereas intrinsic preference cannot change in the same way. Contextual or intrinsic preference may change due to changing beliefs. A person may prefer to pursue an academic career, but adjust his preference if he learns about the consequences of being a professor.

Finally, preference often comes with a ceteris paribus proviso, which refers to the condition of "other things being equal". Therefore, in preference logic special attention is given to ceteris paribus preference. Moreover, to formalize the use of value and preference in practical reasoning, the logic of preference needs to be developed further, for instance, combined with the logic of belief.

## 3 Data-driven Ethical Decision-making

With the rise of complex AI systems and the advancements of their autonomy, the issues of ethical decision-making is receiving growing attention since, as it the case with human decision makers, these systems can be confronted with critical situations in which the decisions to be made can have heavy ethical consequences. These type of situations are also knows as moral dilemmas in which, for example, a self-driving car needs to take a decision in a critical situation where human life is at stake. Traditionally, the normative systems and deontological logics governing the behavior of these proposed systems relied on complex ethical principles and policies that should be formally specified [16,8]. Nevertheless, the problem of endowing AI system with the capacity to make ethical decisions has remain a challenging tasks for the past years [44]. This issue is further complicated by the fact that ethical principles are often dynamic, across cultures, geographies, as well as other human related factors (e.g., gender). These differences were elegantly highlighted by the Moral Machine experiment [7].

The recent breakthroughs in big-data and machine learning made their way into AI ethics and novel data-driven approaches have shown remarkable preliminary results promising to tackle practical solutions. In this context, data are not only used evaluate the outcomes, but are also used to produce new inference models. More specifically, in many cases, it is difficult to rely on formal specifications to reflect the actual decision making situation as implied by the real-data sensed by the agent. Instead, machine learning mechanisms are used to mine implicit reasoning patterns in the data and use them to predict new cases. The quantification and modeling of such norms and values can be performed via machine learning and parametric techniques using Inverse Reinforcement Learning (IRL) [1], imitation learning [42], inverse game theory [45], and norm inference [20].

Despite these recent advances, many challenges persist:

C1 **Data-hungry algorithms**: Most of off-the-shelf machine learning mechanisms require big training datasets. Thus, when applied for Data-driven ethical decision-making, these mechanism would require training datasets to be representative of societal choices and ethical values [32].

C2 **Human-Trained Machine Learning**: Human behavior and performance should provide the baseline to teach the AI and benchmark its behavior against humans. However, most of existing works fail to define and measure human operator performance in real-time context [8]. Moreover, in some applications, value and its ranking are the tacit knowledge possessed by human participants. Contained in the data related to the subject, it is difficult to express in a formal way, and it is also difficult to obtain and process.

C3 **Context-dependence**: Ethical decision-making is mostly context-dependent. Thus, compared to the often brittle traditional approaches, data-driven mechanisms are easier to adapt to dynamically changing scenarios. However, this necessitates that training data are representative of the different contexts.

C4 **Black-box Machine Learning Mechanisms**: most of machine learning mechanisms used for data-driven ethical decision-making are black-box mechanisms whose inner-workings are subsymbolic and, therefore, non-understandable by humans.

## 4 Norm- and Value-Based Multiagent Systems

In dynamic, open and heterogeneous societies, agents are expected to be collaborating with other agents (human and artificial agents). This is the case, for instance, of self-driving cars, medical robots, and home service robots who all demonstrate autonomous capabilities. In these application scenarios, there are often interactions and collaborations between multiple intelligent agents, in the emerging, dynamic, open and multi-agent system, the autonomous behavior and decision-making of each intelligent agent may have a huge impact on society. If left unchecked, self-interested agents my cause harms to others while trying to seek their goals. Similar to their roles in human societies, norms provide a means to regulate agent behavior [27] and make them conform to certain social expectations at the ethical and legal levels. Thus, the results of the face-to-face output conform to certain social expectations at the ethical level, such as: the need to add norms to the system to guide the overall behavior of the system, and to monitor and control the behavior of intelligent agents.

Nevertheless, developing normative multi-agent system raises a specific set of challenges:

C1 **Norms in Open vs Closed Agent Societies**: Often, traditional normative systems are designed for closed societies whose value system becomes the basis for the normative system. The rules stipulated reflect the general importance of society. But when the environment in the society is open, which is the case of most of multi-agent systems, we cannot guarantee that the individual agents entering the society have the same value system as the society. (e.g., when a self-driving car enters a new traffic environment). Detecting and resolving this conflict arising between the two normative systems and ensuring that individual agents have compatible value systems in order to ensure that the overall behavior of the society meets our expectations are open challenges.

C2 **Legalistic vs interactionist view of norms**: The former considers that the normative agent system as a regulatory instrument regulating the emerging behavior of open systems without enforcing the desired behavior. In such a case, agents are often motivated by sanctions to stick to norms, rather than by their sharing of the norms, whereas the later (the interactionist view) is a bottom-up approach in which norms can be seen, as regularities of behavior which emerge without any enforcement system because agents conform to them either because (i) their goals happen to coincide (ii) because they feel themselves as part of the group (iii) or because they share the same values of other agent. In this case, sanctions become not always necessary even for norm violation [11,13].

C3 **Subjective and Cross-cultural differences**: as has been shown by recent user studies [7,21], norms tend to be dependent on factors such as culture (i.e., European vs east Asian), to be context-dependent and application dependent, and to be different from one user to another. In order to cope with this challenge, it is necessary to study the relationship between the subject's decision-making based on the value system and norms and the overall behavior of the system.

## 5 Evaluation in Multi-culture

The existing research on norms and value reasoning is generally based on a specific geographical or cultural background [23,6]. But norms and values greatly differ in and of the different countries, societies, communities, and individuals. This aspect has to be taken account when developing formal systems of reasoning with norms and values.

Law is a geographically-socially determined system of norms: the legal systems—both on the level of (constitutional) values and the very norms based on the former—are different in the different countries. These dissimilarities are especially strong between countries with different histories, cultural and political backgrounds. And while it would be rather difficult to take all countries' all norms and values into account, aiming for a cross-cultural reference and considering some relevant differences between China and the European countries is definitely feasible and needed for a comprehensive approach of an adequate normative reasoning within AI.

One might expect that ethics is less divisive than law, but apparently this is far from sure. The well-known and often discussed Moral Machine[7] is a cross-cultural study of ethical norms and values based on self-driving car scenarios. While there are many critical opinions regarding its methodology, the simulation[8], and the questions used in the experiment, what the study apparently clearly shows is that there are no globally valid ethical considerations, moral norms. The Moral Machine experiment and study concerned a specific environment of self-driving cars with different scenarios realizing variants of the so-called Trolley Problem[14], so one might think that one a higher, or more abstract level, the values and norms might converge. But an investigation of 84 newly written ethical guidelines it was found that while there are some emerging values, "no single ethical principle appeared to be common to the entire corpus of documents".[19] Next to the object level of—specific or abstract—norms, cultural and historical differences might affect the meta-level considerations too: what source should be accepted as a source of valid norms, who can and should decide about what norms should drive the mechanisms of, for example, autonomus systems.

This takes us to different tasks when developing formal models of reasoning with norms and values for new generation AI. One one hand, we need to take these differences into account when considering and collecting the norms and values themselves. There might be differences also in the way people reason with the norms and values in the different cultures and these possible differences have to accommodated too, but first data need to be gained on the existing dissimilarities. On the top of it, we need a resilient reasoning system which, optimally, is not only applicable for different ways of reasoning with norms and values, but also applicable in situations where meta-reasoning is needed about what norms and values should be applied in the reasoning. On the other hand, these differences have to be taken into consideration when verifying the reasoning system: it has to be checked against different benchmark examples coming from different cultures and evaluated in different environments. Collecting data as input for developing an adequate reasoning system and the results of its evaluation against the different cultures and their norms and values will provide an extensive comparative study between China and the European countries.

For taking this aspect seriously, one needs to establish methods for collecting data from different cultures and for evaluating the adaptability of the formal model in a cross-cultural context to answer the question: how does the model perform in different geographical and cultural contexts?

To do so, the we identify the following steps:

(i) Devise new metrics allowing to measure how adaptable a formal model to different cultures, and how data can be collected in an invasive-less manner. A similar endeavor is being conducted in the domain of explainable AI where new metrics are being defined [17] to measure the satisfaction and trustworthiness inspired by an explanation. In addition, a distinction is made between objective and subjective understandability, thereby giving room for personal and perhaps inter-cultural differences with regards to explaination reception and understandability.

(ii) Psychometric scales and statistical significance: when conducting user studies, adequate answer scales should be selected. Moreover, the statistical significance of the outcomes should be tested. One good option is the Likert scale [2]. The latter is commonly used in research and surveys to measure attitude, providing a range of responses to a given question or statement. The typical Likert scale is a 5- or a 7-point ordinal scale used by participants to rate the degree to which they agree or disagree with a question or a statement. While the Likert scale is widely used in scientific research, there has been a long-standing controversy regarding the analysis of ordinal data [41]. In fact, analyzing the outcomes of the Likert scale, and the use of parametric tests to analyze ordinal data in general, has been subject to an active and ongoing debate. In order to adequately obtain the user answers, the psychometric scales to be used should be selected, and the methods used to establish the statistical significance should be updated accordingly.

## 6 Towards Reasoning with Norms and Values in New Generation AIs All Around the Globe

Norms and values drive humans' everyday life and decisions. Since new-generation AIs operate in our society and cohabit with humans, we expect these new generation AIs to take norms and values into account, and their logics have to accommodate reasoning about these norms and values. Despite the recent move to discuss AI from interdisciplinary standpoint covering ethical, legal and societal aspects, most of this engagement originates from Euro-American scholars with obvious influences from the Western epistemic tradition. This results in the marginalization of non-western knowledge systems in the study of AI ethics.

In order to make AI acceptable for global audience, several barriers and centrisms needs to be overcame and a more inclusive approach involving east Asia, but also Africa and the middle-east should be devised. The resulting intercultural approach to the ethics of AI should inform the formation of policies and guidelines to regulate the design and use of AI. This research direction has been recommended by several recent initiatives from IEEE[40] and UNESCO advocating a global instrument on the ethics of AI, which would also serve as guidelines for practitioners, governments and policy-makers.

These differences, profoundly shape the nature of contributions to the field on issues of data privacy, social robotics, conceptions of artificial moral agency, moral status and patiency, autonomous weapons systems, big data and the likes. For instance, as noted by Metz [30,31], there are recurrent salient features that can be found in many sub-Saharan cultures that are not found (in the same way) elsewhere in the world. This does not mean these features cannot be found in other cultures, it just means that they are more recurrent in Africa. The same can be said of Western, Middle Eastern and South-East Asian cultures[38]. Hence, both Afro-ethical (e.g. Ubuntu traditions) and Confucian ethical systems share similarities since both are collectivists systems whose normative principles rest heavily on a collectivist disposition notably when it comes to determine right or wrong action. In contrast to Western ideals, built on advancing individualism following the age of enlightenment and the values spread by the industrial revolution, Afro-ethical and Confucian moral values share principles that advance collective progress, harmony and group cohesion.

Also, to accommodate the clearly existing differences between the different cultures, we have to gather insights also about the tacit or vague norms and values people reason with in order to make the logics used for normative reasoning in artificial intelligence legitimate. In order to do so, next to the usual theoretical methodologies of developing new formal systems, we need to engage with data-driven and experimental methodologies with people, agents and texts, in order to reveal both the norms and values, and the ways people reason with them in different cultures. Employing in this wide range of methodologies realizing a cross-disciplinary endeavor serves the purpose of gaining a comprehensive overview of what reasoning with norms and values is suitable for the new generation of AIs.

## References

[1] Abel, D., J. MacGlashan and M. L. Littman, *Reinforcement learning as a framework for ethical decision making.*, , **16**, Phoenix, AZ, 2016, p. 02.

[2] Albaum, G., *The likert scale revisited*, Market Research Society. Journal. **39** (1997), pp. 1–21.

[3] Alchourrón, C. E. and E. Bulygin, "Normative Systems," Springer-Verlag New York – Wien, 1971.

[4] Anjomshoae, S., A. Najjar, D. Calvaresi and K. Främling, *Explainable agents and robots: Results from a systematic literature review*, in: *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088.

[5] APT, K. R., *Chapter 10 - logic programming*, in: J. VAN LEEUWEN, editor, *Formal Models and Semantics*, Handbook of Theoretical Computer Science, Elsevier, Amsterdam, 1990 pp. 493–574.
URL `https://www.sciencedirect.com/science/article/pii/B9780444880741500159`

[6] Awad, E., M. Anderson, S. L. Anderson and B. Liao, *An approach for combining ethical principles with public opinion to guide public policy*, Artif. Intell. **287** (2020), p. 103349.
URL `https://doi.org/10.1016/j.artint.2020.103349`

[7] Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon and I. Rahwan, *The moral machine experiment*, Nature **563** (2018), pp. 59–64.

[8] Behzadan, V., J. Minton and A. Munir, *Trolleymod v1. 0: An open-source simulation and data-collection platform for ethical decision making in autonomous vehicles*, in: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 391–395.

[9] Bench-Capon, T. and S. Modgil, *Norms and value based reasoning: Justifying compliance and violation*, Artificial Intelligence and Law **25** (2017), pp. 29–64.

[10] Benzmüller, C., X. Parent and L. van der Torre, *Designing normative theories for ethical and legal reasoning: Logikey framework, methodology, and tool support*, Artificial Intelligence **287** (2020), p. 103348.
URL `https://www.sciencedirect.com/science/article/pii/S0004370219301110`

[11] Boella, G., L. Van Der Torre and H. Verhagen, *Ten challenges for normative multiagent systems*, in: *Dagstuhl Seminar Proceedings*, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2008.

[12] Broersen, J., M. Dastani, J. Hulstijn, Z. Huang and L. van der Torre, *The boid architecture - conflicts between beliefs, obligations, intentions and desires*, in: *In Proceedings of the Fifth International Conference on Autonomous Agents* (2001), pp. 9–16.

[13] Castelfranchi, C., *Modelling social action for ai agents*, Artificial intelligence **103** (1998), pp. 157–182.

[14] Foot, P., *The problem of abortion and the doctrine of the double effect*, Oxford Review **5** (1967), pp. 5–15.

[15] Gabbay, D., J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, "Handbook of Deontic Logic and Normative Systems," College Publications, 2013.

[16] Greene, J., F. Rossi, J. Tasioulas, K. Venable and B. Williams, *Embedding ethical principles in collective decision support systems*, , **30**, 2016.

[17] Hoffman, R. R., S. T. Mueller, G. Klein and J. Litman, *Metrics for explainable ai: Challenges and prospects*, arXiv preprint arXiv:1812.04608 (2018).

[18] Horty, J. F., "Reasons as Defaults," Oxford University Press, 2012.

[19] Jobin, A., M. Ienca and E. Vayena, *The Global Landscape of AI Ethics Guidelines.*, Nature Machine Intelligence **1** (2019), pp. 389–399.

[20] Kasenberg, D., T. Arnold and M. Scheutz, *Norms, rewards, and the intentional stance: Comparing machine learning approaches to ethical training*, in: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 184–190.

[21] Komatsu, T., B. F. Malle and M. Scheutz, *Blaming the reluctant robot: parallel blame judgments for robots in moral dilemmas across us and japan*, in: *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 63–72.

[22] Langley, P., B. Meadows, M. Sridharan and D. Choi, *Explainable agency for intelligent autonomous systems.*, , **17**, 2017, pp. 4762–4763.

[23] Liao, B., M. Anderson and S. L. Anderson, *Representation, justification, and explanation in a value-driven agent: an argumentation-based approach*, AI Ethics **1** (2021), pp. 5–19.
URL `https://doi.org/10.1007/s43681-020-00001-8`

[24] Liao, B., N. Oren, L. Van Der Torre and S. Villata, *Prioritized norms in formal argumentation*, Journal of Logic and Computation **29** (2019), pp. 215–240.
URL `https://hal.archives-ouvertes.fr/hal-02381116`

[25] Liao, B., M. Slavkovik and L. van der Torre, *Building Jiminy Cricket: An Architecture for Moral Agreements Among Stakeholders*, in: *Proceedings of the 2019 AAAI/ACM Artificial Intelligence, Ethics and Society* (2019), pp. 147–153.

[26] Luck, M., S. Mahmoud, F. Meneguzzi, M. Kollingbaum, T. J. Norman, N. Criado and M. S. Fagundes, *Normative agents*, in: *Agreement technologies*, Springer, 2013 pp. 209–220.

[27] Luck, M., S. Mahmoud, F. Meneguzzi, M. Kollingbaum, T. J. Norman, N. Criado and M. S. Fagundes, "Normative Agents," Springer Netherlands, Dordrecht, 2013 pp. 209–220.
URL `https://doi.org/10.1007/978-94-007-5583-3_14`

[28] Makinson, D., *On a fundamental problem of deontic logic*, in: *Norms, Logics and Information Systems. New Studies in Deontic Logic and Computer Science*, IOS press, 1999 pp. 29–53.

[29] Makinson, D. and L. van der Torre, *Input/output logics*, Journal of Philosophical Logic **29** (2000), pp. 383–408.

[30] Metz, T., *Toward an african moral theory*, Journal of Political Philosophy **15** (2007), pp. 321–341.

[31] Metz, T., *Toward an african moral theory (revised edition)*, in: *Themes, issues and problems in African philosophy*, Springer, 2017 pp. 97–119.

[32] Noothigattu, R., S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar and A. Procaccia, *A voting-based system for ethical decision making*, , **32**, 2018.

[33] Parent, X. and L. van der Torre, *Input/output logic*, in: D. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, editors, *Handbook of Deontic Logic and Normative Systems*, College Publications, 2013 pp. 353–406.

[34] Rao, A. S. and M. P. Georgeff, *Bdi agents: From theory to practice*, in: *IN PROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON MULTI-AGENT SYSTEMS (ICMAS-95*, 1995, pp. 312–319.

[35] Reiter, R., *A logic for default reasoning*, Artificial Intelligence **13** (1980), p. 81–132.

[36] Russell, S. J. and P. Norvig, *Artificial intelligence: a modern approach* (2010).

[37] Schlechta, K., "Nonmonotonic Logics," Springer, 1997.

[38] Segun, S. T., *Critically engaging the ethics of ai for a global audience*, Ethics and Information Technology (2020), pp. 1–7.

[39] Sergot, M. J., F. Sadri, R. A. Kowalski, F. Kriwaczek, P. Hammond and H. T. Cory, *The british nationality act as a logic program*, Commun. ACM **29** (1986), p. 370–386.
URL `https://doi.org/10.1145/5689.5920`

[40] Shahriari, K. and M. Shahriari, *Ieee standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems*, in: *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, IEEE, 2017, pp. 197–201.

[41] Sullivan, G. M. and A. R. Artino Jr, *Analyzing and interpreting data from likert-type scales*, Journal of graduate medical education **5** (2013), p. 541.

[42] Taylor, J., E. Yudkowsky, P. LaVictoire and A. Critch, *Alignment for advanced machine learning systems*, Ethics of Artificial Intelligence (2016), pp. 342–382.

[43] Von Wright, G., *Deontic logic*, Mind : a Quarterly Review of Psychology and Philosophy **60** (1951).
URL `http://search.proquest.com/docview/1293708393/`

[44] Wallach, W. and C. Allen, "Moral machines: Teaching robots right from wrong," Oxford University Press, 2008.

[45] Wang, Y., Y. Wan and Z. Wang, *Using experimental game theory to transit human values to ethical ai*, arXiv preprint arXiv:1711.05905 (2017).

[46] Wright, G. H. v. G. H., "Norm and Action : a Logical Enquiry," International library of philosophy and scientific method, Routledge and Kegan Paul, London, 1963.