Contesting Black-Box AI Decisions

Blue Sky Ideas Track

Virginia Dignum
Umeå University
Umeå, Sweden
virginia.dignum@umu.se

Loizos Michael
Open University of Cyprus &
CYENS Center of Excellence
Nicosia, Cyprus
loizos@ouc.ac.cy

Juan Carlos Nieves
Umeå University
Umeå, Sweden
jcnieves@cs.umu.se

Marija Slavkovik
University of Bergen
Bergen, Norway
marija.slavkovik@uib.no

Julliett Suarez
University of Granada
Granada, Spain
julliettsuarez@correo.ugr.es

Andreas Theodorou
Universitat Politècnica de Catalunya
Barcelona, Spain
a.theodorou@bath.edu

ABSTRACT

The “right to contest” decisions that have consequences on individuals or the society is a well-established democratic right. Contesting a decision is not a matter of simply providing an explanation, but rather of assessing whether the decision and the explanation are permissible against an organization’s governance framework. Yet, albeit the popularity of adjacent fields, little work has been explicitly done on contesting AI decisions. In this paper, we propose that formal argumentation can be used to formulate contestations of decisions made by artificial agents. We extend the discourse on socio-ethical values in AI by conceptualizing our argumentation framework as a formal dialogue, enabling the interaction between humans and agents as decisions are being contested.

KEYWORDS

explainable AI, contestable AI, formal argumentation, algorithmic decision making

ACM Reference Format:
Virginia Dignum, Loizos Michael, Juan Carlos Nieves, Marija Slavkovik, Julliett Suarez, and Andreas Theodorou. 2025. Contesting Black-Box AI Decisions: Blue Sky Ideas Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 5 pages.

1 INTRODUCTION

The increased use of artificial intelligence (AI) in decision making, by both public and private institutions, may increase efficiency in decision making, but there is a risk of making the reasoning of those decisions unclear and out of reach for direct contestation by the beneficiaries of the institution. The interactions between individuals and institutions occur through well-defined channels and protocols; however, they are still direct person-to-person interactions even when people do not meet directly face-to-face. If formal interactions do not run smoothly, face-to-face is the option of last recourse when we seek clarifications and resolutions. What if the institution substitutes their human representatives with an artificial agent or an automated artificial intelligence system? Who can a beneficiary face? How can the volume of automated decisions be matched by the real need of beneficiaries to understand and contest them? All these emergent issues can become difficult or impossible to trace back to their source, contributing to an *accountability gap*.

The existing legal framework, such as Article 21 of EU’s General Data Protection Regulation (GDPR) [11], guarantees the right to object to the processing of personal data and to challenge decisions. To operationalize these rights in a digital context, their integration into the system’s design so as to effectively handle disputes is crucial [3]. When decisions are made by algorithms rather than humans, clarifying what constitutes a contestable decision becomes essential, involving identifying key entities and their interrelations in the contestation process. The emerging field of *contestable AI*, focusing on enabling AI systems to accommodate interventions throughout their operational life-cycle, addresses these needs [4, 14, 21].

This paper discusses the need to design a framework to facilitate contestations against the decisions of opaque information systems. We analyze an institutional decision-making process dependent on an AI model for decision-making, assuming a lack of transparency in the model?s operations. We provide a possible conceptual framework for contestation. We emphasize the need for AI systems not only to provide explanations that clarify their internal decision-making processes, but also to offer justifications that validate these decisions as sound and acceptable, aligning them with external norms and standards to ensure robust accountability.

Contributions. The main contribution of this paper is the introduction of a novel feedback architecture to monitor AI-driven predictions and decisions against specified norms, legal requirements, and ethical standards. The proposed framework allows for evaluating the contestability of black-box decisions, enhancing oversight and assessment capabilities for decisions by autonomous systems. We promote the development of AI systems that are ethical, transparent, and contestable, fostering increased trust, accountability, and societal responsibility and compliance of AI technologies.

This paper is structured as follows: In the next section, we provide an overview of related work to the topic of constable AI. In Section 3 we present our conceptual architecture for contestability of black-box decisions. In Section 4 we introduce a formalization of

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. NowÃľ (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

contestability based on formal argumentation. Lastly, in Section 5 we present our conclusions and outline directions for future work.

2 BACKGROUND AND RELATED WORK

A decisions can only be contested on some grounds that relate to the process of how it has been reached. Therefore, we first look towards the, now relatively rich, field of explainable Artificial Intelligence.

Justifications, not Explanations. Explainable Artificial Intelligence (xAI) is a growing field that seeks to provide insights into how algorithms arrive at their decisions and recommendations [6]. Explanations in AI transfer knowledge from automated systems to aid human comprehension of why decisions are made [10, 19].

Despite the value of explanations, there is a concern that attention has deviated from the need to provide *justifications* [13]. Justifications aim to affirm that a decision is sound or acceptable, while explanations clarify the processes behind these decisions. Establishing robust justifications is crucial for accountability, addressing both the intrinsic and extrinsic criteria that automated decision systems must meet [7, 8]. Justifications typically relate to the internal or external norms of the organizations that deploy these systems, standing apart from the algorithms themselves.

Article 22 of EU's General Data Protection Regulation (GDPR) mandates that individuals must have the right to contest decisions made solely by automated processes, emphasizing the need for robust justification mechanisms [12]. However, the GDPR does not specify the exact procedures for contesting these decisions or whether justifications should be provided by the automated systems or their deploying organizations. Nonetheless, it is clear that the entity responsible must justify the appropriateness of any decision made against specified requirements, not merely explain it.

The distinction between justifications and explanations has also been debated [16]. While they are both forms of reasoning, they serve different roles: justifications prove the validity of a conclusion, while explanations describe the rationale behind it. This complexity arises because the same facts might serve both to justify and explain, and the language used in both contexts — such as 'because', 'so', 'therefore' — overlaps. Ultimately, whether an argument is seen as a justification or an explanation can depend on the intent behind its presentation. In fact, one can view justifications as explanations that are *compliant* against an AI system's decision-making specifications, which in turn could be the norms, laws, and standards that determine the context in which the AI system is contested [24].

Contestability in the Literature. Kluttz et al. [18] define contestability as mechanisms for users to understand and challenge model predictions. Lyons et al. [21] see it as a post-decision process involving explanations and user contests. These definitions focus on the system itself being able to provide the means for contesting its decisions. Taking a more socio-technical view, Aler Tubella et al. [1] emphasize the need for explicit values elicitation and alignment between the system and the rational behind a decision. Alfrink et al. [3] view it as a tool for human interaction with AI systems throughout the operational life-cycle of those systems.

Within the wider human-AI interaction field, contestability has been seen as a *design principle.* Alfrink et al. [3] argue that systems designed with contestability in mind can enable their users to detect and correct errors and record disagreements with decisions made [14]. Almada [4] considers public perception and advocates early integration of contestability into system design. Kluttz et al. [17] treat contestability as a system's design principle; achievable by providing feedback mechanisms such as expert users critiquing and correcting the system?s reasoning. Providing feedback to correct a system's errors has been the focus of multiple user studies and frameworks on interactive machine learning [5, 23, 25, 28].

Formal Argumentation Theory. The process of contesting can be seen as a dialogue between the institution in charge of the AI system and an affected agent by the outcome of the AI system. In each interaction of a contesting process, the agents can provide arguments of different nature, e.g., explanations, justification, facts, etc. [20]. Formal argumentation theory has explored different kinds of dialogues, e.g., negotiation, information-seeking, etc. [29]. Nevertheless, there are no dialogue frameworks that could deal with the interactive process of a contesting process, although [20] argue that computational argumentation is ideally suited for this task. Most of the works regarding the use of argumentation on black-box systems aim to build explanations on the top of the outcome of the black-box system [9] or assess the compliance of a black-box system with respect to some given social norms [2].

Some of the research gaps to characterize a contesting process as a dialogue process are to identify which information should be contained in the arguments that are posted during a contesting process, as well as to identify and resolve the conflicts between the arguments that are provided during the contesting process. Having clear definitions of these issues can provide means to characterize different processes for contesting on black-box AI systems.

Deontic statements. Weigand and Dignum [30] consider communication as the origin of norms, and explore how deontic statements are created and adapted in communication processes. They consider a normative system to be a set of interacting agents. The authors propose a formal language that specifies certain speech acts and the deontic statements that they bring about. For example, if an employer has authority to demand for a certain task to be done by an employee, asking for the task creates an obligation for the employee to execute it. Our contestation system can be seen as a normative system in the sense of [30], if we broaden the understanding of speech acts to include the interlocutions between an AI system and a human affected by the decisions of the former.

The interplay between argumentation and deontic statements (such as obligations and permissions) has been considered in the literature. van der Torre and Villata [27] consider the problem of enriching legal argumentation with a formal account of deontic modalities. Specifically, they propose a formal framework that allows one to reason over normative concepts and to compare norms. Although their framework does not explicitly consider processes of contestation, it can still be used to specify an argumentation framework that can generate obligations and permissions.

3 CONCEPTUAL FRAMEWORK

In this section, we present our conceptual framework. To this end, explicit definitions of different actors and entities are introduced. Algorithmic accountability establishes a relationship between an agent (i.e. actor) and a forum, where the agent must explain and justify algorithmic behaviors to a forum that has the authority to question and pass judgment [31]. This relationship extends to include various roles: *Representatives:* Agents responsible for modifications to the algorithm. *Beneficiaries:* Individuals directly impacted by the algorithm, possessing rights to appeal the algorithm's decisions. *Third Parties:* Entities authorized to audit algorithmic operations.

Contestability is both a characteristic of a system that facilitates these interactions, as described by Alfrink et al. [3], and a process that obliges representatives to account for algorithmic actions to beneficiaries and third parties, as argued by Aler Tubella et al. [1]. We adopt both perspectives, but we focus on the operational implications of contestability. Operationalizing contestability may involve either local or global modifications to the algorithm through appeals or audits, respectively. The outcome of this process often results in an obligation for representatives to amend or justify algorithmic behaviors. Setting clear obligations is the first step in ensuring the compliance with requirements for providing justifications, and taking measures when compliance is not met.

Our approach, emphasizing the integration of contestability into AI systems, ensures that automated decisions are not 'just' transparent, but also can be subject to scrutiny, having to demonstrate the validity and permissibility of their decisions and their underlying reasoning, and as such facilitate accountability. This perspective is crucial whenever decisions made by AI can significantly impact individuals and communities. By enabling systematic appeals and requiring justifications that align with socio-ethical values and legal mandates, our framework ensures that AI systems adhere to higher standards of fairness and justice. This proactive incorporation of contestability helps prevent harm, builds public trust, and ensures that AI systems are aligned with human rights and democratic values, making them more sustainable and acceptable in society.

We assume the algorithm operates as a black box, observable only through its inputs and outputs. This assumption ensures that our framework is applicable to any algorithm whose behavior can be observed, irrespective of its internal workings. The black box concept is extended to include the organization utilizing the algorithm, highlighting the broader context of algorithmic accountability.

To contest a black box system, a beneficiary must be able to initiate a process that compels the system's representatives to justify the algorithm's compliance with specific criteria. This involves assessing the algorithm against agreed-upon standards or preferences. The output of the system is used to trigger the appeal process, where beneficiaries can review and contest specific decisions, prompting a detailed justification from the system's representatives that must align with established ethical and regulatory guidelines.

Contestability differs from explainability, as it demands not only clear reasons in support of decisions, but also requires the reasons to be justified against certain requirements. For instance, an explanation for a loan denial may be that the decision took into consideration the age of the applicant. Such an explanation may help understand why the decision was made, but it does not constitute a valid justification under anti-discrimination laws. Instead, further information to demonstrate how the decision adheres to the institutional policy are needed; e.g., the applicant should be of legal age, and the loan should end prior to applicant's retirement.

Generating such justifications during a contestation process considers three elements: *1. Input:* Contesting data provided to the black box, such as data accuracy or relevance. *2. Algorithm:* Challenging the black box's overall fairness, alignment with values and policy, accuracy, or consistency. *3. Output:* Questioning the adequacy or correctness of the decision outcomes. A specific input, the algorithm used — as a black box — to map that input into an output, and the output itself, together comprise the *case elements* of an appeal. Along with the case elements, an appeal also determines the representative of the algorithm, the specific beneficiary affected by the algorithmic decision, along with the context and the basis for contestation. The appeal can challenge any of the case elements.

4 OPERATIONAL FRAMEWORK

In this section, we introduce an operationalization of contestability, focusing less on covering all aspects presented in Section 3, and more on the role of formal argumentation as a central component of the proposed operationalization. We forego a detailed presentation of the technical details — which can be found in [24] — and elaborate, instead, upon the key ideas behind our operationalization.

We operationalize contestability as a dialogue between a beneficiary and a representative, each taking turns to introduce a new argument in the discussion. Various types of arguments are meant to capture different forms of reasons offered by the parties involved in the contesting process; by the representative, on the one hand, in support of the decision taken by an AI system, and by the beneficiary, on the other hand, in objection to the alignment of that decision and its supporting reasons with agreed-upon requirements:

- Arguments based on values. E.g., The decision is not fair.
- Arguments based on norms. E.g., Other people in my income bracket have been approved for a loan just recently.
- Arguments based on factual errors. E.g., You claimed that I am not employed, but here is my employment record.
- Arguments based on instantiation errors. E.g., You operationalised the norm incorrectly. You selected a wrong norm.
- Arguments based on misplaced counts-as. E.g., You assumed that my children are dependent, but they are adults.
- Arguments based on similarity. E.g., You answered affirmatively to a person who has the same properties as me.
- Arguments based on counter examples. E.g., You said this was the only way to do X, but here is another way to do it.

This list of arguments is illustrative rather than exhaustive. What types of arguments are allowed is defined by the context of the decision, and also by the policy and legal norms that govern the institution. An automatic or computer-aided arbitration of contesting would require an institution to set a strict and well-defined list of the kinds of arguments and evidence that are allowed [15]. This list could, however, be itself a subject of contestation and audits.

Abstracting away from the specifics of the arguments, we can represent each argument as a pair of a premise and a conclusion, both defined over a language (e.g., a certain fragment of first-order logic, over a particular set of predicates) agreed upon by the representative and the beneficiary as part of the contestation context.

The first argument is put forward by the representative in support of the decision made by the AI system. In the simplest case, such an argument can state that its conclusion (i.e. the system's decision to be supported) is supposed to be true. Such *supposition* arguments (e.g., stating that "Your loan should be rejected.") are not meant to ground their conclusion on some evidence, but they, rather, act as placeholders for the "open fronts" of the conversation, so that not everything needs to be fully justified up front. Accordingly, supposition arguments are easy to dispute, simply by having the opposing party suppose the negation of their conclusion (e.g., stating that "My loan should not be rejected."). This dispute is, in effect, equivalent to the opposing party asking a "why" question, forcing the party that initially put forward a supposition argument to identify a different argument to justify that same conclusion.

In the general case, the argument put forward in support of a conclusion (be it the system's decision, or any other intermediate claim) associates the truth of its conclusion to the truth of its premises. Such *association* arguments (e.g., stating that "You cannot apply for a youth loan, because you are over 18.") ground their conclusion on evidence coming from other properties that are relevant to the context of the contestation process. The opposing party has two ways to dispute an association argument. First, the opposing party can present another argument that supports a conflicting conclusion (e.g., stating that "I can apply for a youth loan, because I am less than 25 and unemployed.") that is stronger than the former argument. Whether an argument is stronger than another can be determined by external factors (e.g., an explicit exception in the conditions for being granted a youth loan), or it could, itself, be a matter of dispute and contestation. Second, the opposing party can dispute the premises of the association argument. Instead of introducing a second mechanism to do that, we simply require that whenever an association argument is put forward, its premises are supported by supposition arguments. Thus, the opposing party can simply dispute those supposition arguments, effectively "opening up" the conversation on some other intermediate claim.

Naturally, arguments eventually need to be grounded on perceived facts. Such *perception* arguments (e.g., stating that "Your ID card shows that you are 26 years old.") ground their conclusions on the inputs that the AI system used to reach its decision. Even these arguments can, however, be disputed under certain circumstances, if, for example, the AI system's input includes factual errors, or is perceived incorrectly. As for association arguments, argument strength ultimately determines which disputes succeed and which do not. In general, however, supposition arguments are the weakest, and perception arguments are the strongest, leaving the remaining relative strengths to be determined by the contestation context.

Through this exchange of arguments, we end up with a sequence of argumentation frameworks that alternate between entailing and not entailing the system's decision. When either the representative or the beneficiary fails to extend this alternating sequence by providing strong arguments in their favor, this terminates the *contesting dialogue* by, respectively, *upholding* or *rejecting* the appeal.

The set of all contesting dialogues that can be realized for certain case elements and in a particular contestation context can be depicted as a *deliberation argument tree*, whose vertices correspond to (sets of) arguments, whose edges correspond to disputes between (sets of) arguments with conflicting conclusions, whose root corresponds to the system's decision that is being contested, and whose branches correspond to individual contesting dialogues. A deliberation argument tree can be extended to associate each of its leaves with the obligations and permissions to be generated as a result of upholding or rejecting the appeal in the corresponding branch.

5 DISCUSSION AND CONCLUSIONS

The democratic right to challenge decisions affecting individuals or the society is firmly established, yet, unexplored within the wider AI literature. Contesting a decision entails more than providing a mere explanation; it involves evaluating whether the decision and its supporting reasons align with externally provided policies, even if it may not be the most favorable choice for the decision maker. In this paper, we proposed a framework to enable the contestation of decisions made by AI systems; thereby improving oversight and evaluation of autonomous system decisions.

Our architecture uses argumentation logic to generate, given a policy, justifications for a decision made by a system in a structured, and inherently transparent, process for contesting decisions. Our framework can be used for the automatic verification of a decision against a given normative policy. This verification can be done post-hoc, but also as part of a monitoring system in a feedback loop; similar to a human-in-the-loop approach, our contestation framework can verify and approve each decision made. Ultimately, by enabling the contestation of AI decisions, our framework promotes accountability as we ensure that decisions accepted and contested by the various actors are subject to audit and verification. The operationalization of contestability described herein has been further formalized [24], and implemented as part of a prototype arbitration system [15]. Similarly, [22] presents a formalisation for making a contestable reasoner for generating context-specific explanations.

It has been argued that a certain way to evaluate the moral bounds of an AI system, is by creating a 'glass box', i.e. a monitoring mechanism of the system's inputs and outputs that is able to run tests to check the system's compliance against some values [26]. To construct such a glass box one needs to be able to translate values into design requirements, and then demonstrate that it effectively establishes necessary obligations and permissions, promoting transparency. Our contesting framework based on formal argumentation supports the interaction between different stakeholders exchanging various arguments. The contestation process outcomes help us setup a normative framework for the actors in the system.

By enhancing the contestability of black-box AI systems, this work paves the way for more responsible and sustainable AI applications across various sectors, and sets the foundation for contestable AI work in policy studies, normative systems, and formal argumentation. While this initiative is a step toward building trust in AI, the realistic impact on achieving fully responsible and sustainable AI implementations across diverse applications will require ongoing development and empirical validation. Within policy studies, future work needs to identify the scaffolding of the types of rights and affordances that should be allowed in a contestation process in general and in specific contexts. Within normative systems, future work needs to establish how these rights and affordances are to be encoded into norms and obligations that are generated and validated as part of the contestation process. Lastly, within formal argumentation, future work needs to elaborate the types of supported arguments to allow for obligations, permissions and other norms to be instated as a result of a successful appeal process.

ACKNOWLEDGMENTS

This work was funded by HORIZON project HumanE-AI-Net Grant agreement ID: 952026 as micro-project "Contesting Black-Box Decisions".

REFERENCES

[1] Andrea Aler Tubella, Andreas Theodorou, Virginia Dignum, and Loizos Michael. 2020. Contestable Black Boxes. In *Rules and Reasoning*, Víctor Gutiérrez-Basulto, Tomáš Kliegr, Ahmet Soylu, Martin Giese, and Dumitru Roman (Eds.). Springer International Publishing, Cham, 159–167.

[2] Andrea Aler Tubella, Andreas Theodorou, and Juan Carlos Nieves. 2021. Interrogating the Black Box: Transparency through Information-Seeking Dialogues. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 106–114.

[3] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2022. Contestable AI by Design: Towards a Framework. *Minds and Machines* (Aug. 2022). https://doi.org/10.1007/s11023-022-09611-z

[4] Marco Almada. 2019. Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (Montreal, QC, Canada) *(ICAIL '19)*. Association for Computing Machinery, New York, NY, USA, 2âĂŞ11. https://doi.org/10.1145/3322640.3326699

[5] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (Dec. 2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[6] Alejandro Barredo Arrieta, Natalia DÃŋaz-RodrÃŋguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador GarcÃŋa, Sergio Gil-LÃşpez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. http://arxiv.org/abs/1910.10045 Number: arXiv:1910.10045 arXiv:1910.10045 [cs].

[7] Reuben Binns. 2018. Algorithmic Accountability and Public Reason. *Philosophy & Technology* 31 (12 2018), 1–14. https://doi.org/10.1007/s13347-017-0263-5

[8] Claude Castelluccia and Daniel Le MÃľtayer. 2019. Understanding algorithmic decision-making: Opportunities and challenges. Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit within the Directorate-General for Parliamentary Research Services (DG EPRS) of the Secretariat of the European Parliament.

[9] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, Francesca Toni, et al. 2021. Argumentative XAI: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. 4392–4399.

[10] "European Commission High-Level Expert Group on AI, Ethics guidelines for trustworthy AI" 2019.

[11] European Parlament & Council. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). http://data.europa.eu/eli/reg/2016/679/oj.

[12] European Parliament and Council of European Union. 2016. General Data Protection Regulations (GDPR). Place: EU.

[13] Talia B Gillis and Josh Simons. 2019. Explanation < Justification: GDPR and the Perils of Privacy. Issue 71. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3374668

[14] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E. Imel, and David C. Atkins. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 conference on designing interactive systems (DIS '17)*. Association for Computing Machinery, New York, NY, USA, 95–99. https://doi.org/10.1145/3064663.3064703 Number of pages: 5 Place: Edinburgh, United Kingdom.

[15] Christodoulos Ioannou and Loizos Michael. 2024. An Automated Arbitrator for Contesting Dialogues. In *Proceedings of the International Workshop on AI Value Engineering and AI Compliance Mechanisms (VECOMP 2024)*.

[16] Tziporah Karachkoff. 1988. Explaining and Justifying. *Informal Logic* 10 (1988). https://doi.org/10.22329/il.v10i1.2635

[17] Daniel N. Kluttz, Nitin Kohli, and Deirdre K. Mulligan. 2020. Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. In *After the digital tornado: Networks, algorithms, humanity*, KevinEditor Werbach (Ed.). Cambridge University Press, Cambridge, 137–152.

[18] Daniel N Kluttz, Deirdre K Mulligan, Kenneth Bamberger, Solon Barocas, Michael Buckland, Jenna Burrell, Tim Casey, Taylor Cruz, Madeleine Elish, Fernando Delgado, James X Dempsey, Anne-Laure Fayard, Anna Lauren Hoffman, Karrie Karahalios, Nitin Kohli, Karen Levy, Clifford Lynch, Chris Mammen, Jonathan Marshall, Susan Nevelow Mart, Scott Skinner-Thompson, and Jennifer Urban. 2019. Automated Decisions Support Technologies and the Legal Profession. *Berkeley Technology Law Journal* 34 (2019), 853. https://doi.org/10.15779/Z38154DP7K Concept of contestability.

[19] Christian Lahusen, Martino Maggetti, and Marija Slavkovik. 2024. Trust, trustworthiness and AI governance. *Scientific Reports* 14, 1 (Sept. 2024), 20752. https://doi.org/10.1038/s41598-024-71761-0

[20] Francesco Leofante, Hamed Ayoobi, Adam Dejl, Gabriel Freedman, Deniz Gorur, Junqi Jiang, Guilherme Paulino-Passos, Antonio Rago, Anna Rapberger, Fabrizio Russo, Xiang Yin, Dekai Zhang, and Francesca Toni. 2024. Contestable AI Needs Computational Argumentation. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*. 888–896. https://doi.org/10.24963/kr.2024/83

[21] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. *Proceedings of the ACM on Human-Computer Interaction* 5 (4 2021). Issue CSCW1. https://doi.org/10.1145/3449180

[22] Leila Methnani, Virginia Dignum, and Andreas Theodorou. 2024. Clash of the Explainers: Argumentation for Context-Appropriate Explanations. In *ECAI 2023 International Workshop: XAI^3*, Vol. 1947. Springer Nature Switzerland, Cham, 7–23. https://doi.org/10.1007/978-3-031-50396-2_1 Series Title: Communications in Computer and Information Science.

[23] Loizos Michael. 2023. Autodidactic and Coachable Neural Architectures. In *Compendium of Neurosymbolic Artificial Intelligence*. IOS Press, 235–248. https://doi.org/10.3233/FAIA230143

[24] Loizos Michael. 2024. Explanatory Compliance. In *Proceedings of the 5th Workshop on Explainable Logic-Based Knowledge Representation (XLoKR 2024)*.

[25] Stefano Teso and Kristian Kersting. 2019. Explanatory Interactive Machine Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Honolulu HI USA, 239–245. https://doi.org/10.1145/3306618.3314293

[26] Andrea Aler Tubella, Andreas Theodorou, Frank Dignum, and Virginia Dignum. 2019. Governance by Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 5787–5793. https://doi.org/10.24963/IJCAI.2019/802

[27] Leendert van der Torre and Serena Villata. 2014. An ASPIC-based legal argumentation framework for deontic reasoning. *Frontiers in Artificial Intelligence and Applications* 266: Computational Models of Argument (2014). https://doi.org/10.3233/978-1-61499-436-7-421

[28] Jesse Vig, Shilad Sen, and John Riedl. 2011. Navigating the tag genome. In *Proceedings of the 16th international conference on Intelligent user interfaces*. ACM, Palo Alto CA USA, 93–102. https://doi.org/10.1145/1943403.1943418

[29] Douglas Walton and Erik CW Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.

[30] Hans Weigand and Frank Dignum. 1994. *Deontic Aspects of Communication*. John Wiley & Sons, Inc., USA, 259âĂŞ273.

[31] Maranke Wieringa. 2020. What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT*20)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3351095.3372833