can verify consistency between a policy and the mental states of a suspect agent. The methods for determining whether the suspect agent is consistent with the policy effectively provide an acceptance policy: if the suspect agent's assertions are consistent with the policy for a given definition of consistency, then it is determined that the agent satisfactorily adheres to the policy.

Definition 4.21 (Interrogation). Let $Ag^S = \langle P, \hat{f}_{\mathcal{D}}, d, g \rangle$ be a suspect agent, $Ag^I = \langle P, \hat{f}_D, q \rangle$ be an investigator agent and σ be an argumentation semantics.

- Ag^I strongly believes in Ag^S if for all $Topic \in q(P)$, Ag^S σ-sceptically argues about $Topic$ and $P \cup Topic_P \nvdash \bot$.
- Ag^I credulously believes in Ag^S if for all $Topic \in q(P)$, Ag^S σ-credulously argues about $Topic$ and $P \cup Topic_P \nvdash \bot$.
- Ag^I strongly does not believe in Ag^S if for all $Topic \in q(P)$, Ag^S σ-empty argues about $Topic$ and $P \cup Topic_P \nvdash \bot$.

We can choose to determine that policy is being followed when Ag^I strongly believes in Ag^S, or when Ag^I credulously believes in Ag^S, depending on how strict we wish to be. These acceptance protocols are by no means exclusive: acceptance can be made to depend on other aspects of the argumentation semantics, or even other aspects of the dialogue itself.

The framework we have presented offers many possibilities due to the modularity of its components. We can tune several aspects of i) query generation, ii) aggregation of knowledge obtained from the system being inspected, and iii) acceptance criteria for compliance. In addition, the output of this process is a dialogue, containing topics, assertions about this topic, as well as arguments extracted from this dialogue and whether these arguments are accepted. All of this output is directly human-readable, and provides a level of transparency about both the functioning of the learning system being inspected and the compliance checking process itself.

5 EXAMPLE IMPLEMENTATION

In this section, we describe an implementation of the framework for a small subset of the dataset of the running example described in Section 3. Our aim is to study the adherence of this recommender system to the policy P described in Example 4.4. Our recommender system is trained on the whole dataset, but for brevity of exposition we limit our queries to a small subset of it. We need to emphasise to the reader that our contribution is not the recommender system, but rather the framework used to evaluate it.

The recommender system combines two popular approaches, *content based* [1] and *collaborative Filtering* [9], into a hybrid system that first finds similar movies as the one inputted by the user and then ranks them based on the users' profile. For the movie search, we use movies' metadata such as cast, crew, genre, and keywords to calculate the cosine similarity between movies. Once calculated, we take the 20 closest movies and then rank them using our collaborative filtering approach. Collaborative Filtering is based on the notion that users similar to other users would rate items the same way. We use Singular Value Decomposition (SVD) algorithm to create our model. SVD has been made popular since its use by the winning team of the Netflix Grand Prize winner [21, 22]. Our SVD model predicts the ratings a user would give to the 20 movies selected by our content-based part of the system. The system then ranks them based on those predictions and present only the top 10 to the user. We selected these techniques mentioned below due to their robustness, speed, and commonality in movie recommender systems.

Table 1: Dialogue between the Investigator Agent and Suspect Agent for the topic (X_w, V).

Name	Move
m_1	$\langle Ag^I, open, (X_w, V) \rangle$
m_2	$\langle Ag^S, assert, \langle \mathbf{x_1}, \texttt{highVariety} \rangle \rangle$
m_3	$\langle Ag^S, assert, \langle \mathbf{x_2}, \texttt{highVariety} \rangle \rangle$
m_4	$\langle Ag^S, assert, \langle \mathbf{x_3}, \texttt{mediumVariety} \rangle \rangle$
	$\vdots$
m_{12}	$\langle Ag^S, close \rangle$
m_{13}	$\langle Ag^I, close \rangle$

The formal dialogue framework for this recommender system is given by the investigator agent and suspect agent described in Examples 4.16 and 4.18. We aim to study whether the investigator agent strongly/credulously believes in the suspect agent with relationship to the policy P. This depends on the existence of a well-formed dialogue for every $Topic \in q(P)$. As described in Example 4.16, there are seven topics produced by the investigator agent. For each of these topics, such a dialogue is produced by having the suspect agent assert all of the black box arguments related to the topic as produced by the argument generator g. For example, for the topic (X_w, V), the produced dialogue is shown in Table 1.

From this dialogue γ, we extract the arguments and their attack relations. In this case, arguments attack each other when their support is similar (same user, similar movie), but the descriptor is different. We therefore obtain an argumentation graph $AF_\gamma = \langle A_\gamma, Att(A_\gamma) \rangle$ where:

- $A_\gamma = \{1, \ldots, 10\}$
- $Att(A_\gamma) = \{(2,8), (8,2), (2,9), (9,2), (2,6), (6,2), (2,3), (3,2), (3,6), (6,3), (3,8), (8,3), (3,2), (2,3), (4,6), (6,4), (6,3), (3,6), (6,8), (8,6), (6,4), (4,6), (6,2), (2,6), (8,6), (6,8), (8,3), (3,8), (8,10), (10,8), (9,2), (2,9), (10,8), (8,10)\}$

We are denoting arguments by the number of the move on which the argument was presented by the suspect agent. Let us apply two classical argumentation semantics the so-called grounded and stable semantics [15] to AF_γ[2]: the results are shown in Table 2.

From these results, we can observe that the recommender system sceptically argues about the arguments $\{1, 5, 7\}$. Hence, the recommender system has strong beliefs on arguments such as $1 := \langle \mathbf{x_1}, \texttt{highVariety} \rangle$. However, there are arguments such as $2 := \langle \mathbf{x_2}, \texttt{highVariety} \rangle$ that is low represented in the stable extensions. Hence, the investigator agent can believe that the recommender system has low evidence about Argument 2. By using Definition 4.21 and the results of $\sigma_{ground}(AF_\gamma)$ and $\sigma_{stable}(AF_\gamma)$, the investigator agent can verify the compliance of different policies. Let us observe that the grounded and stable semantics are only two argumentation semantics from a big variety of argumentation

[2]We used the argumentation solver: http://gerd.dbai.tuwien.ac.at/.

Table 2: Extensions for the argumentation graph AF_γ for grounded and stable semantics.

Argumentation semantics	Extensions
$\sigma_{ground}(AF_\gamma)$	$\{1, 5, 7\}$
$\sigma_{stable}(AF_\gamma)$	$\{9, 8, 7, 5, 4, 1\}$
	$\{10, 9, 7, 6, 5, 1\}$
	$\{10, 7, 5, 4, 2, 1\}$
	$\{10, 9, 7, 5, 4, 3, 1\}$

semantics that exits in the state of the art of formal argumentation reasoning [5]. Hence, the selection of a proper argumentation semantics for implementing an investigator agent can be a question on its own.

6 RELATED WORK

The perspective of testing whether a learning agent complies with a policy in fact sets this work within the general area of *conformance testing*. Conformance testing approaches for "black box" and adaptive systems are still being developed: a specific challenge is that of the breadth of the test space [13]. The framework proposed in this paper is related to a breadth of literature on agents testing other agents, particularly those approaches which propose to construct an agent or a multi-agent system with the explicit purpose of testing another agent. For example, [26] propose to construct an agent that can generate tests from the ontologies describing a MAS being tested, after which responses to these tests are verified. Similar in outlook, [32] propose a framework consisting of a multi-agent system made-up of a testing agent, a monitoring agent and agents representing the task environment, particularly focused on identifying goals that are not being met by the agent being tested.

In the sense that our framework produces an argumentation graph modelling the behaviour of the learning agent being inspected, our approach is also related to work on *agents modelling other agents*. The literature is vast in this topic, in the context of multi-agent systems especially, given that in collaborative or competitive scenario it is often needed to produce a model of other agents to predict what their behaviour will be [12, 33]. Formal argumentation methods are more often used to model communication between agents [29] and agent knowledge [6], but there are approaches that use argumentation frameworks to build an *opponent model* representing what another agent believes based on a dialogue [18, 30, 36]. These are similar in outlook to the framework we present: in a way, we are representing the learning agent's beliefs in the form of properties that hold for its input/output pairs, in what resembles building a machine theory of mind [28].

7 CONCLUSIONS AND FUTURE WORK

In this paper we present a modular framework for evaluating a learning system's adherence to a policy. The formal dialogue framework we present is based on the idea of building an argumentation framework representing the arguments expressed in a dialogue between an investigator agent and a suspect agent. In this way, we construct a model of the learning agent by considering its properties across inputs. A strength of this approach is given by the modularity of its components, each of which can be implemented in a variety of ways depending on which properties of the learning system we wish to study. Additionally, the use of a dialogue as an information-seeking tool provides a level of transparency about the querying and testing process. Finally, we propose acceptance criteria determining adherence to a policy that are fundamentally different from the quantitative approaches often used: acceptance is determined through argumentation semantics, suggesting a new notion of "consistent compliance" to a policy. The limits of this approach lay where access to the learning system is limited: when it is not possible to construct inputs matching the policy or to describe outputs with the predicates given by the policy.

Future work is planned on several directions, exploiting both the versatility of the framework and the potential for studying specific logical properties of learning agents. An important development is to extend this to more sophisticated representations of policies, with more sophisticated languages allowing for capturing the complexity of social and ethical norms. Beyond that, a refinement of this framework would be to study the possibilities of topic generation and output aggregation. For example, a possibility is to exploit the topic generator function of the investigator agent to adversarially generate those topics that are expected to yield more inconsistencies in the dialogue. Another is to implement the description function of the suspect agents (describing inputs/outputs in terms of the policy) as a learning system itself, which learns which are good or bad outputs in terms of the policy. An exciting possibility is to implement a description function that aggregates several inputs, returning descriptions of the output together with a probability weight: this would provide a hybrid quantitative/qualitative approach to aggregating knowledge about a learning system.

A further interesting research avenue is to study the presentation of the "degree of agreement" to a policy, in a way that is most useful to foster trust and promote transparency. In this paper we have proposed two possible degrees– sceptical and credulous– but many other possibilities exist. Additionally, we aim to exploit methods developed for this framework to study technical properties of learning agents, such as monotonicity or rationality, similarly to how we defined the notion of non-monotonic belief checking.

Overall, we believe this framework offers distinct benefits in terms of modularity and transparency, as well as opening the door to new non-quantitative ideas of compliance. Furthermore, it opens many possibilities in terms of future development for the purpose of better understanding, and controlling, the learning systems that are becoming increasingly pervasive in our society.

ACKNOWLEDGMENTS

A. Aler Tubella was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Theodorou A. is funded by the Knut and Alice Wallenberg Foundation, grant agreement 2020.0221.

REFERENCES

[1] Charu C. Aggarwal. 2016. *Recommender Systems.* Springer International Publishing. https://doi.org/10.1007/978-3-319-29659-3

[2] Huib Aldewereld, Virginia Dignum, and Yao Hua Tan. 2015. Design for values in software development. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*. Springer Netherlands, 831–845. https://doi.org/10.1007/978-94-007-6970-0{_}26

[3] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society* 20, 3 (3 2018), 973–989. https://doi.org/10.1177/1461444816676645

[4] S Anjomshoae, A Najjar, D Calvaresi, and K Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)* (2019). http://www.diva-portal.orghttp//urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-158024

[5] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. 2011. An introduction to argumentation semantics. *Knowledge Eng. Review* 26, 4 (2011), 365–410.

[6] Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review* 33, 3 (3 2010), 211–259. https://doi.org/10.1007/s10462-010-9154-1

[7] Or Biran and Courtenay Cotton. 2017. Explanation and Justification in Machine Learning: A Survey. *IJCAI Workshop on Explainable AI (XAI)* August (2017), 8–14. https://pdfs.semanticscholar.org/02e2/e79a77d8aabc1af1900ac80ceebac20abde4.pdf

[8] Elizabeth Black and Anthony Hunter. 2009. An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems* 19, 2 (2009), 173–209.

[9] John S. Breese, David Heckerman, and Carl Kadie. 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI'98)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 43–52.

[10] Joanna J Bryson and Andreas Theodorou. 2019. How Society Can Maintain Human-Centric Artificial Intelligence. In *Human-Centered Digitalization and Services*, Marja Toivonen-Noro, Evelina Saari, Helinä Melkas, and Mervin Hasu (Eds.). Springer, 305–323. https://doi.org/10.1007/978-981-13-7725-9{_}16

[11] Joanna J. Bryson and Alan Winfield. 2017. Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. *Computer* 50, 5 (5 2017), 116–119. https://doi.org/10.1109/MC.2017.154

[12] David Carmel and Shaul Markovitch. 1996. Opponent modeling in multi-agent systems. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 1042. Springer Verlag, 41–52. https://doi.org/10.1007/3-540-60923-7{_}18

[13] Camille Constant, Thierry Jéron, Hervé Marchand, and Vlad Rusu. 2007. Integrating formal verification and conformance testing for reactive systems. *IEEE Transactions on Software Engineering* 33, 8 (8 2007), 558–574. https://doi.org/10.1109/TSE.2007.70707

[14] Virginia Dignum. 2019. *Responsible artificial intelligence : how to develop and use AI in a responsible way*. Springer.

[15] Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence* 77, 2 (1995), 321–358.

[16] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human Computer Studies* 58, 6 (2003), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

[17] Xiuyi Fan and Francesca Toni. 2014. A general framework for sound assumption-based argumentation dialogues. *Artif. Intell.* 216 (2014), 20–54. https://doi.org/10.1016/j.artint.2014.06.001

[18] Christos Hadjinikolis, Yiannis Siantos, Sanjay Modgil, Elizabeth Black, and Peter Mcburney. [n.d.]. *Opponent Modelling in Persuasion Dialogues.* Technical Report.

[19] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

[20] Taemie Kim and Pamela Hinds. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication* (2006), 80–85. https://doi.org/10.1109/ROMAN.2006.314398

[21] Y. Koren. 2009. The BellKor Solution to the Netflix Grand Prize.

[22] Yehuda Koren and Robert Bell. 2011. *Advances in Collaborative Filtering*. Springer US, 145–186. https://doi.org/10.1007/978-0-387-85820-3_5

[23] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (1 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50{_}30392

[24] Joseph E. Mercado, Michael A. Rupp, Jessie Y.C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors* 58, 3 (5 2016), 401–415. https://doi.org/10.1177/0018720815621206

[25] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *AI and Society* 1 (2 2020), 3. https://doi.org/10.1007/s00146-020-00950-y

[26] Cu Duy Nguyen, Anna Perini Fondazione, Bruno Kessler, Paolo Tonella Fondazione, Cu D Nguyen, and Anna Perini. 2008. Experimental Evaluation of Ontology-Based Test Generation for Multi-agent Systems. (2008). https://doi.org/10.1007/978-3-642-01338-6{_}14

[27] Simon Parsons, Michael Wooldridge, and Leila Amgoud. 2003. Properties and Complexity of Some Formal Inter-agent Dialogues. *Journal of Logic and Computation* 13, 3 (2003), 347–376.

[28] Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine Theory of Mind. *35th International Conference on Machine Learning, ICML 2018* 10 (2 2018), 6723–6738. http://arxiv.org/abs/1802.07740

[29] Chris Reed and Doug Walton. 2005. Towards a formal and implemented model of argumentation schemes in agent communication. *Autonomous Agents and Multi-Agent Systems* 11, 2 (9 2005), 173–188. https://doi.org/10.1007/s10458-005-1729-x

[30] Tjitze Rienstra, Matthias Thimm, and Nir Oren. 2013. Opponent Models with Uncertainty for Strategic Argumentation. , 332–338 pages. https://abdn.pure.elsevier.com/en/publications/opponent-models-with-uncertainty-for-strategic-argumentation

[31] Tracy L. Sanders, Tarita Wixon, K. Elizabeth Schafer, Jessie Y. C. Chen, and P. A. Hancock. 2014. The influence of modality and transparency on trust in human-robot interaction. In *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. IEEE, 156–159. https://doi.org/10.1109/CogSIMA.2014.6816556

[32] Francisca Raquel Vasconcels Silveira, Gustavo Augusto Lima Campos, and Mariela Ines Cortes. 2013. Rational agents for the test of rational agents. *IEEE Latin America Transactions* 11, 1 (2013), 651–657. https://doi.org/10.1109/TLA.2013.6502879

[33] Dicky Suryadi and Piotr J. Gmytrasiewicz. 1999. Learning models of other agents using influence diagrams. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 407. Springer Verlag, 223–232. https://doi.org/10.1007/978-3-7091-2490-1{_}22

[34] Andreas Theodorou and Virginia Dignum. 2020. Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence* 2, 1 (1 2020), 10–12. https://doi.org/10.1038/s42256-019-0136-y

[35] Andreas Theodorou, Robert H. Wortham, and Joanna J. Bryson. 2017. Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science* 29, 3 (2017), 230–241. https://doi.org/10.1080/09540091.2017.1310182

[36] Matthias Thimm. 2014. Strategic Argumentation in Multi-Agent Systems. *KI - Kunstliche Intelligenz* 28, 3 (8 2014), 159–168. https://doi.org/10.1007/s13218-014-0307-2

[37] Elliot Turiel. 2001. *The Culture of Morality*. Cambridge University Press. https://doi.org/10.1017/cbo9780511613500

[38] Douglas Walton and Erik C. W. Krabbe. 1995. Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning. https://philpapers.org/rec/WALCID

[39] Alan F.T. Winfield and Marina Jirotka. 2017. The case for an ethical black box. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10454 LNAI. 262–273. https://doi.org/10.1007/978-3-319-64107-2{_}21

[40] Ryan W. Wohleber, Kimberly Stowers, Jessie Y.C. Chen, and Michael Barnes. 2017. Effects of agent transparency and communication framing on human-agent teaming. In *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017*, Vol. 2017-Janua. IEEE, 3427–3432. https://doi.org/10.1109/SMC.2017.8123160