Interspeech 2024
1-5 September 2024, Kos, Greece

# On the Use of *Plausible Arguments* in Explainable Conversational AI

*Martina Di Bratto*[1,3], *Maria Di Maro*[2,3], *Antonio Origlia*[2,3]

[1]Dept. of Humanities, University of Naples Federico II, Italy
[2]Dept. of Electrical Engineering and Information Technology, University of Naples Federico II, Italy
[3]Urban/Eco Research Center, Italy

martina.dibratto@unina.it, maria.dimaro2@unina.it, antonio.origlia@unina.it

## Abstract

Conversational Artificial Intelligence has evolved to facilitate more efficient communication of user preferences through dialogues. This paper delves into Argumentative Conversational AI systems, and, more specifically in the definition of a methodology of selecting and using *plausible arguments* to support recommendations. We propose a cross-disciplinary model grounded in cognitive pragmatics to enhance recommendation quality. We evaluate this linguistically motivated strategy in isolation using simulated dialogues and collect human judgements to verify that the expected interaction is believable. Next, we test the full interaction model with human users to evaluate its usability. Results indicate high scores for naturalness and argument selection, validating the system's plausibility and effectiveness. Concerning usability, the system is perceived as attractive and reliable although technical issues concerning the system's reactiveness are present.

**Index Terms**: Argumentation, Conversational AI, Plausibility, Explainability

## 1. Introduction

This work presents a dialogue system architecture based on a linguistically motivated conversation management model equipped with argumentation capabilities. The model is tested in the domain of movie recommendations and draws inspiration from a set of cognitive properties for data representation mapped on corresponding network analysis measures. The proposed system leverages on a graph-based representation of the domain combined with dynamic information collected during the dialogue to form a belief system. Beliefs concerning the user opinions towards domain items, together with the measures describing the items' role in the domain, form the basis upon which arguments are selected to support the system's stances (i.e., recommendations). The architecture separates the interaction model from a knowledge graph representation of the domain, to build an explainable Conversational Artificial Intelligence (AI). This approach aims at developing a general computational framework for argumentation-based dialogue which is missing [1], although dialogue systems with argumentative skills have been developed for specific tasks [2, 3, 4]. The theoretical aspects of argumentation have been investigated, mostly from an understanding point of view, e.g. considering annotation schema [5] and automatic detection of argument components [6, 7, 8] (for a full survey, see [9]).

From a theoretical point of view, argumentation entails verbal, social, and rational processes designed to convince a reasonable critic of the validity of a viewpoint. This is achieved by presenting a *constellation* of propositions, constituting an illocutionary complex speech act, either supporting or challenging the proposition inherent in the viewpoint [10]. The selection of the most appropriate items and features is pivotal for the achievement of the conversational goal. In this work, a theoretical model for argument selection will be described and evaluated through its computational implementation. Differently from a perspective purely based on machine learning, our proposal leverages on the mathematical interpretation of a set of cognitive properties for *plausible* arguments. Current research efforts are concentrating on introducing graph-based knowledge to support Large Language Models (LLM), especially in the field of Retrieval Augmented Generation (RAG) [11, 12]. This approach aims at selecting relevant information from which an LLM should extract an answer to a question. However, from a goal-oriented point of view, this approach still does not consider the system's *objectives* in the conversation. LLMs, in this setup, have the role of natural language generation modules with no decision making functionalities, which are implemented using other kinds of model. This *hybrid* architecture aims at combining the best of symbolic and sub-symbolic representations to equip Conversational AIs with different sets of capabilities, orchestrated by a dedicated model.

## 2. Background theory

In this work, we propose a theoretical framework to select arguments based on *plausibility* and effectiveness of the system's argumentation strategies in guiding users towards optimal choices. Concerning plausibility, however, it does not guarantee the truth or correctness of an argument as a plausible argument may still be incorrect or incomplete. Conversely, plausibility is a subjective assessment that depends on the evaluation of the dialogue participants or the audience and can be defined as the degree of coherence or effectiveness of an argument within the dialogue. It assesses how relevant and believable a new argument is, based on the quality of the new data and also on its connection with data already available. Specifically, a plausible argument is one that appears to be well-supported, logical, and consistent with the available information and common ground [13, 14]. In our proposal, we map the theoretical concepts of data plausibility onto numerical aspects computed over the graph database structure. More specifically, being the implementation of our dialogue management module based on a graph database hosting a knowledge base of common facts collected from Linked Open Data sources [15], methods typical of network analysis were used to represent, in line with theory, the correct selection of arguments. The theoretical motivations for the possibility to use the theoretical concepts adopted in our proposal within the knowledge structure is summarised in Figure 1 and explained below. According to [14]'s data networks, there are two cases of argumentation through plausibility, shown in


10.21437/Interspeech.2024-839

Figure 1a: i) self-evident data having a large number of data connections that support them; ii) explanatory data which, in turn, are connected to many other data to support them. This work highlights the strong connection between explainability and argument plausibility on which our . At the same time, from a mathematical point of view, a datum is considered *authoritative* among the others when it has several data supporting it (i.e, self-evident datum), whereas a datum has a high *hub* score when it supports other kind of authoritative data (i.e., explanatory data), as shown in Figure 1b. Hubs and authorities represent what [16] called a *mutually reinforcing relationship* meaning that a good hub is a node that points to many good authorities, while a good authority is a node that is referred to by many good hubs. This graph representation of data allows us, by recovering the alternative model of belief revision (Data-oriented belief revision, DBR) [17], to retrieve concepts of cognitive pragmatics describing the quality of an argument and to mathematically retrieve this same information within a graph structure. DBR is, therefore, used to determine the subset of reliable information (i.e., beliefs) and their degree of strength. In their model data are selected as beliefs on the basis of their properties, i.e., the possible cognitive reasons to believe such data: *credibility, importance, relevance and (un-)likeability*. An important aspect for the selection of this model is the measurability of the features, as our system translates them into numerical descriptors of the graph structure, supporting explainability.

Specifically, **Credibility**, defined as a measure of the number and values of all supporting data, contrasted with all conflicting data, down to external and internal sources, is mapped on the nodes' **Authority** score, describing the importance of their role in the graph since a solid number of nodes with a high hub score support their validity. **Importance**, a measure of the epistemic connectivity of the datum (i.e., the number and values of the data that the agent will have to revise, should they revise that single one) is mapped on **Hub** scores, describing the nodes' connection degree to authoritative nodes. **Relevance**, defined as a measure of the pragmatic utility of the datum (i.e., the number and values of the pursued goals that depends on that datum) is mapped on **Entropy**, indicating about which nodes the user's stance is most uncertain, so that obtaining feedback about them is important. **(Un-)likability** is mapped on **Hard evidence**, representing the system's beliefs.Leveraging the above-described measures, the system selects the most appropriate arguments to recommend, and using them or their features to support its decision. The mapping between cognitive properties and their mathematical interpretation constitutes the basis from which the presented computational model has been built.

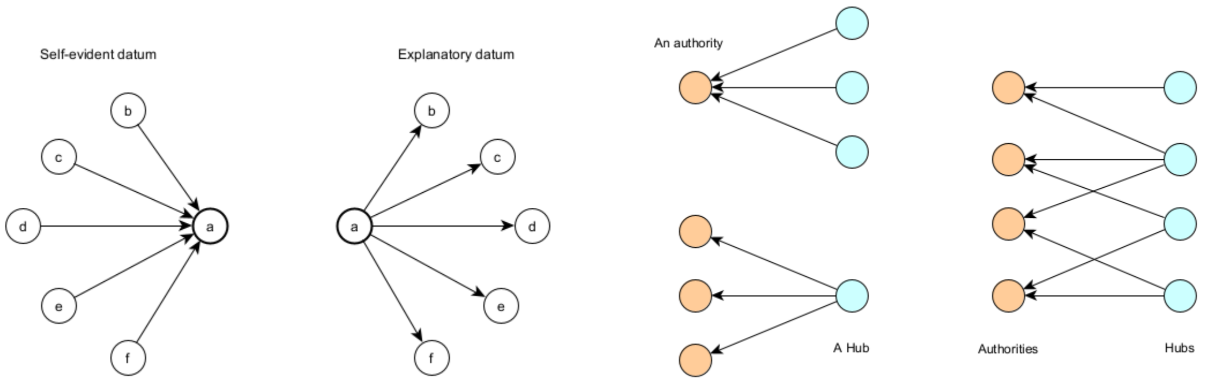
## 3. Technological framework

The presented approach is based on the system architecture assumed by Framework for Advanced Natural Tools and Applications with Social Interactive Agents (FANTASIA) [18, 19]. FANTASIA[1] is a plugin for the Unreal Engine designed to support the development of Embodied Conversational Agents. A FANTASIA Conversational AI follows these main principles: **Behaviour Trees** (BT) [20] are used to organise and prioritise dialogue moves; **Graph Databases** (i.e., Neo4j [21]) are used for knowledge representation and dialogue state tracking; **Bayesian Networks**, implemented using the aGRuM library [22], are used for decision making; LLMs are used to verbalise the decisions taken by PGMs. The proposed interaction model is designed to extract relevant sub-graphs from the knowledge database, analyse their structure, and compute the utility of possible dialogue moves. This produces an explainable-by-design system using machine learning for perception and generation tasks while using human-interpretable models for knowledge representation and decision making. To evaluate the argumentation model, the recommendation task has been chosen due to its inner dialogical structure, which can be divided into *Exploration* and *Exploitation* phases, and the goal it encompasses [23]. The need to select supporting arguments for a recommendation constitutes a secondary goal pursued during the exploration phase, while building the common ground [24, 25, 26]. To satisfy the goal of both types of dialogues, the selection of the most appropriate *items* and *features* is pivotal. Our approach separates the interaction management model, in the form of BTs and PGMs, from knowledge representation in Graph Databases, instantiating different problems. This way, the structure of the BT and the logic for assembling PGMs do not change, among tasks, aiming to represent an *abstract* interaction model. For the scope of this paper, we concentrated on the movie recommendation task, using the knowledge graph described in [15]. When the interaction starts, nothing is known about the user. The system, then, uses data coming from Movielens [27] and the results of the network analysis to extract the most informative sub-graph. To this aim, an extraction query considers multiple aspects to build a ranking of items to be recommended (movies). First of all, as the goal is to recommend a movie that is not obscure nor too well known, a scoring function peaking at maximum $k$ value of 1 at 50% probability while producing 0 values at both 0% and 100% probability levels, scores movies:

$$k = 1 - \left(\frac{2n}{max(n)} - 1\right)^2 \qquad (1)$$

where $n$ is the number of opinions expressed by Movielens users. The utility value of each candidate item is obtained by considering the best balance between the probability of an item to be known and the probability of the same item to be liked. We therefore define a utility function $U_{sel}$ as the harmonic mean of $k$ and the average Movielens opinion score $o$, normalised between 0 and 1. The best three items are taken as reference to build the BN together with their features, like actors, directors and genres. Also, *secondary* items sharing features with the primary ones are considered, according to the following procedure to assemble the BN: a) extract the features of primary items; b) for each primary item, extract the top five most useful secondary items, ranked using $U_{sel}$; c) extract all ontological *part_of* relationships involved in the set of nodes composed by the union of both features and primary/secondary items; d) rank the list of relationships using the authority score of target nodes as the primary sorting value and the hub score of source nodes as secondary sorting value. Ordering candidate relationships first by authority and then by hub scores follows the principles described in the previous Section. Authority is a measure of *credibility*, so that nodes with high support are preferred in the selection. Hub scores measure *importance*, so that, when there are not enough authoritative nodes in the network, the possibility of discovering them is increased by considering nodes with a supporting role. Concerning prior distributions, we apply uniform distributions to nodes with no incoming relationships (actors and directors). All other nodes are represented as aggregator nodes computing the median of their parents. Given the obtained Bayesian Network, ratings distributions extracted from Movielens are applied as soft evidence, computed using

[1] github.com/antori82/FANTASIA




(a) *Source: [14]* (b) *Source: [16]*

Figure 1: *The concepts of plausible arguments (a) mapped on a densely linked set of hub (cyan) and authoritative nodes (orange) (b).*

Kernel Density Estimation to compensate cases with low number of ratings. Beliefs about the user are applied as hard evidence. At this point, it is possible to estimate the utility of recommending each item using $U_{sel}$ with the updated distributions after Bayesian inference. To establish whether to recommend the most useful item, we use a dynamic threshold, implemented as a sigmoid function taking the number of turns as a parameter. The threshold dynamically decreases as the dialogue becomes longer, simulating the necessity to reach a conclusion more *urgently* as the dialogue continues. When a recommendation cannot be provided, the system computes the most useful question to ask. To perform this *exploration* move, the system follows, in order of priority, four different strategies, employing the Markov blanket of the most useful item to recommend the $m$ movie. Collecting user feedback about authoritative nodes is given priority, with higher priority given to authoritative nodes in the Markov blanket of $m$. Secondarily, hub nodes are considered in the same way. Exploration moves can be presented in two forms: an open question or a polar question. The system prefers the former case when the accumulated utility for a certain node *class* (i.e., genres, actors, etc. . . ) accumulates more that 60% of the total nodes utility and the relative size of the class domain is lower that 10% of the total domain size of the candidate items. This leads the system to ask open questions when most of the utility falls in a specific class and the number of options is not large. When the user answers, the graph structure is updated to represent a belief graph according to the feedback and a new set of primary items, consistent with the new beliefs, is extracted together with their features. When an *exploitation* move is performed, supporting features are presented with the suggestion to implement argumentation. Consistently with what has been described in Section 2, the number of features involved in positive beliefs represents the *(un-)likeability* $l$ parameter. The relevance parameter $r$ is the normalised feature entropy of the probability distribution over ratings, after Bayesian inference. The credibility parameter $c$ is the authority score of the feature. The importance parameter $i$ is its hub score. After normalisation, the harmonic mean of the four parameters is considered as a ranking score $U_f$. The first three features in the ranking are used to support the recommendation.

**Algorithm 1** System move generation pseudocode

$updateBeliefs(user, utterance)$
$primNodes \leftarrow$ best three recommendable items
$feats \leftarrow$ features of primary items
$secNodes \leftarrow$ best five secondary items for each primary item
$bn \leftarrow bayesianNetwork(primNodes, secNodes, features)$
$bn \leftarrow inference(softEvidence, hardEvidence, bn)$
$sigThreshold \leftarrow updateThreshold(Turn)$
$recItem \leftarrow bestNode(primNodes, bn)$

**if** $U_{sel}(recItem) >= sigThreshold$ **then**
  $bestFeats \leftarrow$ best three features for $recItem$
  $recommend(recItem, bestFeats)$
**else if** $accumulatesUtility(class)$ **then**
  $queryOpen(class, user)$
**else if** Unexplored authoritative $node$ exists in $MB(primNodes)$ **then**
  $queryPolar(node, user)$
**else if** Unexplored authoritative node $n$ exists in $primNodes$ **then**
…
**end if**

## 4. Methodology

To conduct an initial evaluation of the described hybrid model, we implemented a dialogue simulator to generate synthetic exchanges among a recommender and a seeker in the movies domain. The simulator is composed by the conversational recommender AI, implemented on the basis of the previously described principles, and a batch of simulated users, acting on the basis of a probabilistic model that approximates the expected behaviour of real users using Movielens data and a set of predefined rules. The dialogues for this first experiment aimed to compare simulated dialogues with positive and negative extremes. Specifically, we used: a) **Positive sub-set of the control group**, i.e., five dialogues from the INSPIRED Corpus [28] to represent ideal human-human interactions; b) **Negative subset of the control group**, i.e., five dialogues generated with our system, where both the selection of the target items and of the supporting features during exploitation was randomised; c) **Target group**, i.e., ten simulated dialogues produced using the proposed computational model. For the test, 20 human judges were recruited on the Prolific platform and the test was implemented in Qualtrics. They were not informed that the dialogues were mostly synthetically generated and were asked to provide a score on a Likert scale ranging from 1 (not at all) to 5 (absolutely) for the following questions, designed to let us evaluate the *plausibility* of synthetic dialogues considering also the qual-

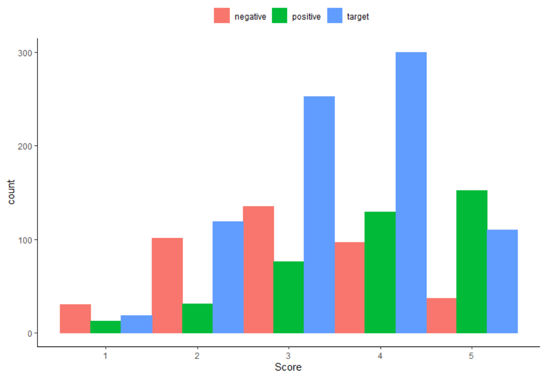



Figure 2: *The full distribution of scores.*

ity of the supporting features: **Q1** Is Mary asking coherent questions to help George finding a movie?; **Q2** Are the two people communicating naturally?; **Q3** Does Mary show a good expertise about the movie domain?; **Q4** Assuming he has not seen the movie, do you think George would accept Mary's suggestion?

All 20 participants evaluated the 20 dialogues in the test, producing 400 data points for each question and a total number of 1600 data points used in the following statistical analysis. The simulator is designed to test the dialogue management strategy independently from other communication problems, linked to speech-based communication. To test the full model, implemented as a FANTASIA-based dialogue management architecture, a dedicated user study has been conducted. A group of 10 people was recruited to interact with the system so that the quality of the argumentative dialogue management model could be evaluated on the field. Participants were left free to interact with the system as long as they wanted until they found an acceptable recommendation that fit their liking. After the interaction, the participants were also asked to compile a questionnaire designed to evaluate technological systems usability: the User Experience Questionnaire (UEQ) [29], which evaluates systems' Attractiveness, Pragmatic, and Hedonic Quality.

## 5. Results

To analyse the data concerning the synthetic dialogues, we employ a Cumulative Link Mixed Model (CLMM) [30] with Laplace approximation, [31]. This model accommodates random effects attributable to individual participants or specific stimuli, treating them as blocking variables and assesses the likelihood of observing high values on the Likert score in relation to the independent variable (i.e., dialogue type). We use one model to analyse the full set of scores and four distinct models to focus on specific questions. The full model detected a highly significant association between positive dialogues and high scores ($p < 0.0001$) while, on the other hand, the negative control dialogues did not exhibit any statistically significant association with high scores. The target group was weakly associated with high scores but the $p$-value was very close to the strong significance threshold ($p = 0.0144$). In Figure 2, the full scores distribution is shown. This model confirms the quality of the collected data, as the participants were able to separate positive from negative samples in the control group. Table 1, summarises the statistical analysis for each question, highlighting the association between high scores and the type of dialogue. While average scores, for target dialogues, outperform the negative baseline, our statistical analysis provides a deeper interpretation. This highlights two aspects: a weak effect found on **Q2** indicates that target dialogues are perceived as more natural than the negative baseline and a strong effect found on **Q4** indicates that target dialogues contain more plausible arguments than the baseline, our main interest in this simulation.

| | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| **Positive** | weak | very strong | very strong | very strong |
| **Negative** | absent | absent | absent | absent |
| **Target** | absent | weak | absent | strong |

Table 1: *Statistical significance of the association between high scores and dialogue types for each considered question.*

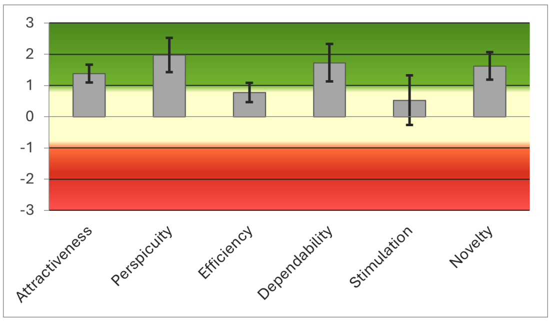



Figure 3: *Participants' ratings about the interaction quality*

The dialogues between the system and the participants lasted, on average, 17 turns. People answered the questions posed by the system and only one participant interrupted the interaction after 24 turns by reporting that the recommendation was not good. The system received positive ratings for Attractiveness, Perspicuity, Dependability, and Novelty with the UEQ. Efficiency and Stimulation received lower scores, mainly due to the system being generally perceived as slow (Figure 3). The main cause of the latency is due to the need of synthesising speech and generating lip-sync animations on the fly. The obtained results are compatible with the estimates obtained with the simulator and demonstrate that our argument selection methodology can be used to implement a linguistically motivated Argumentative Conversational Agent.

## 6. Conclusions

We have presented a model for argumentation dialogues grounded on cognitive principles and on the selection of plausible arguments. Systems based on such principles for argumentation are intrinsically explainable since users are given reason for the machine's behaviour and developers can inspect the decision processes. We have also provided mathematical descriptors for the pragmatic features used in the theoretical model. Using these scores, we have designed a Conversational AI with the capability of using plausible arguments independently of the domain. We have tested the strategy using simulated dialogues and also evaluated the usability of a complete system. Results show that simulated dialogues were scored higher than the negative ones, especially for **Q2** and **Q4**. Moreover users perceived the system as likeable, easy to use, controllable, and innovative. Our findings contribute to advancing research on argumentation applied to AI systems. Future work will consist of deploying the model in different domains and extending its communication capabilities using linguistic theories.

# 7. References


[1] H. Prakken, *Historical overview of formal argumentation*. College Publications, 2018, vol. 1.

[2] H. Nguyen and J. Masthoff, "Designing persuasive dialogue systems: Using argumentation with care," in *Persuasive Technology: Third International Conference, PERSUASIVE 2008, Oulu, Finland, June 4-6, 2008. Proceedings 3*. Springer, 2008, pp. 201–212.

[3] A. R. Panisson, F. Meneguzzi, R. Vieira, and R. H. Bordini, "Towards practical argumentation-based dialogues in multi-agent systems," in *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, vol. 2. IEEE, 2015, pp. 151–158.

[4] W. A. Abro, A. Aicher, N. Rach, S. Ultes, W. Minker, and G. Qi, "Natural language understanding for argumentative dialogue systems in the opinion building domain," *Knowledge-Based Systems*, vol. 242, p. 108318, 2022.

[5] S. Teufel, A. Siddharthan, and C. Batchelor, "Towards discipline-independent argumentative zoning," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 3-EMNLP'09*. Association for Computational Linguistics, 2009.

[6] R. Mochales Palau and M.-F. Moens, "Automatic detection of arguments in legal texts," in *19th Belgian-Dutch Conference on Artificial Intelligence, Date: 2007/11/05-2007/11/06, Location: Utrecht, The Netherlands*, 2007.

[7] T. Goudas, C. Louizos, G. Petasis, and V. Karkaletsis, "Argument extraction from news, blogs, and social media," in *Artificial Intelligence: Methods and Applications: 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May 15-17, 2014. Proceedings 8*. Springer, 2014, pp. 287–299.

[8] M.-F. Moens, "Argumentation mining: How can a machine acquire common sense and world knowledge?" *Argument & Computation*, vol. 9, no. 1, pp. 1–14, 2018.

[9] J. Lawrence and C. Reed, "Argument mining: A survey," *Computational Linguistics*, vol. 45, no. 4, pp. 765–818, 2020.

[10] F. H. Van Eemeren and R. Grootendorst, "A pragma-dialectical procedure for a critical discussion," *Argumentation*, vol. 17, no. 4, pp. 365–386, 2003.

[11] P. Sen, S. Mavadia, and A. Saffari, "Knowledge graph-augmented language models for complex question answering," in *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, B. Dalvi Mishra, G. Durrett, P. Jansen, D. Neves Ribeiro, and J. Wei, Eds. Toronto, Canada: Association for Computational Linguistics, Jun. 2023, pp. 1–8. [Online]. Available: https://aclanthology.org/2023.nlrse-1.1

[12] Q. Yang, Q. Chen, W. Wang, B. Hu, and M. Zhang, "Enhancing multi-modal multi-hop question answering via structured knowledge and unified retrieval-generation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5223–5234.

[13] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on Socially Shared Cognition*, L. B. Resnick, J. M. Levine, and S. D. Teasley, Eds. Washington, DC, USA: American Psychological Association, 1991, pp. 222–233.

[14] F. Paglieri and C. Castelfranchi, "Argumentation and data-oriented belief revision: On the two-sided nature of epistemic change," in *CMNA IV: 4th workshop on computational models of natural argument*. sn, 2004, pp. 5–12.

[15] A. Origlia, M. Di Bratto, M. Di Maro, and S. Mennella, "A multi-source graph representation of the movie domain for recommendation dialogues analysis," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 1297–1306.

[16] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[17] F. Paglieri and C. Castelfranchi, "Revising beliefs through arguments: Bridging the gap between argumentation and belief revision in mas," in *Argumentation in Multi-Agent Systems: First International Workshop, ArgMAS 2004, New York, NY, USA, July 19, 2004, Revised Selected and Invited Papers 1*. Springer, 2005, pp. 78–94.

[18] A. Origlia, F. Cutugno, A. Rodà, P. Cosi, and C. Zmarich, "Fantasia: a framework for advanced natural tools and applications in social, interactive approaches," *Multimedia Tools and Applications*, vol. 78, pp. 13 613–13 648, 2019.

[19] A. Origlia, M. Di Bratto, M. Di Maro, and S. Mennella, "Developing embodied conversational agents in the unreal engine: The fantasia plugin," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6950–6951.

[20] G. Flórez-Puga, M. A. Gomez-Martin, P. P. Gomez-Martin, B. Díaz-Agudo, and P. A. Gonzalez-Calero, "Query-enabled behavior trees," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 1, no. 4, pp. 298–308, 2009.

[21] J. Webber, "A programmatic introduction to neo4j," in *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*. ACM, 2012, pp. 217–218.

[22] G. Ducamp, C. Gonzales, and P.-H. Wuillemin, "aGrUM/pyAgrum : a Toolbox to Build Models and Algorithms for Probabilistic Graphical Models in Python," in *10th International Conference on Probabilistic Graphical Models*, ser. Proceedings of Machine Learning Research, vol. 138, Skørping, Denmark, Sep. 2020, pp. 609–612. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03135721

[23] C. Gao, W. Lei, X. He, M. de Rijke, and T.-S. Chua, "Advances and challenges in conversational recommender systems: A survey," *AI Open*, vol. 2, pp. 100–126, 2021.

[24] H. H. Clark, *Using Language*. Cambridge, UK: Cambridge University Press, 1996.

[25] R. Stalnaker, "Common ground," *Linguistics and philosophy*, vol. 25, no. 5/6, pp. 701–721, 2002.

[26] H. H. Clark and E. F. Schaefer, "Contributing to discourse," *Cognitive science*, vol. 13, no. 2, pp. 259–294, 1989.

[27] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *Acm transactions on interactive intelligent systems (tiis)*, vol. 5, no. 4, pp. 1–19, 2015.

[28] S. A. Hayati, D. Kang, Q. Zhu, W. Shi, and Z. Yu, "Inspired: Toward sociable recommendation dialog systems," *arXiv preprint arXiv:2009.14306*, 2020.

[29] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *Symposium of the Austrian HCI and usability engineering group*. Springer, 2008, pp. 63–76.

[30] A. Agresti, *Categorical data analysis*. John Wiley & Sons, 2012, vol. 792.

[31] Z. Shun and P. McCullagh, "Laplace approximation of high dimensional integrals," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 4, pp. 749–760, 1995.