

Computational argumentation: a foundation for human-centric AI

Edited by

Antonis Kakas, Loizos Michael and Emmanuelle Dietz

Published in

Frontiers in Artificial Intelligence



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-5687-0
DOI 10.3389/978-2-8325-5687-0

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Computational argumentation: a foundation for human-centric AI

Topic editors

Antonis Kakas — University of Cyprus, Cyprus

Loizos Michael — Open University of Cyprus, Cyprus

Emmanuelle Dietz — Airbus, Germany

Citation

Kakas, A., Michael, L., Dietz, E., eds. (2024). *Computational argumentation: a foundation for human-centric AI*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-5687-0

Table of contents

04	Editorial: Computational argumentation: a foundation for human-centric AI Emmanuelle Dietz, Antonis Kakas and Loizos Michael
06	Argumentation: A calculus for Human-Centric AI Emmanuelle Dietz, Antonis Kakas and Loizos Michael
26	Argument-based human–AI collaboration for supporting behavior change to improve health Kaan Kilic, Saskia Weck, Timotheus Kampik and Helena Lindgren
47	An argumentation semantics for rational human evaluation of arguments Marcos Cramer and Leendert van der Torre
60	EQRbot: A chatbot delivering EQR argument-based explanations Federico Castagna, Alexandra Garton, Peter McBurney, Simon Parsons, Isabel Sassoon and Elizabeth I. Sklar
76	Achieving descriptive accuracy in explanations via argumentation: The case of probabilistic classifiers Emanuele Albini, Antonio Rago, Pietro Baroni and Francesca Toni
94	Evaluating and selecting arguments in the context of higher order uncertainty Christian Straßer and Lisa Michajlova
116	Sketching the vision of the Web of Debates Antonis Bikakis, Giorgos Flouris, Theodore Patkos and Dimitris Plexousakis
131	Argumentation and explanation in the law Antonino Rotolo and Giovanni Sartor
148	Argument-based inductive logics, with coverage of compromised perception Selmer Bringsjord, Michael Giancola, Naveen Sundar Govindarajulu, John Slowik, James Oswald, Paul Bello and Micah Clark



OPEN ACCESS

EDITED AND REVIEWED BY
Andrea Passerini,
University of Trento, Italy

*CORRESPONDENCE

Emmanuelle Dietz
✉ emmanuelle.dietz@airbus.com
Antonis Kakas
✉ antonis@ucy.ac.cy
Loizos Michael
✉ loizos@ouc.ac.cy

RECEIVED 05 February 2024

ACCEPTED 01 March 2024

PUBLISHED 18 March 2024

CITATION

Dietz E, Kakas A and Michael L (2024) Editorial:
Computational argumentation: a foundation
for human-centric AI.
Front. Artif. Intell. 7:1382426.
doi: 10.3389/frai.2024.1382426

COPYRIGHT

© 2024 Dietz, Kakas and Michael. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Editorial: Computational argumentation: a foundation for human-centric AI

Emmanuelle Dietz^{1*}, Antonis Kakas^{2*} and Loizos Michael^{3*}

¹Airbus Central R&T, Hamburg, Germany, ²University of Cyprus, Nicosia, Cyprus, ³Open University of Cyprus & CYENS Center of Excellence, Nicosia, Cyprus

KEYWORDS

argumentation, human-centric approach, artificial intelligence, formal foundations, learning, reasoning, cognition

Editorial on the Research Topic

Computational argumentation: a foundation for human-centric AI

1 Introduction

What is an appropriate foundation for building Human-centric AI (HCAI) systems? What foundation would allow AI to draw elements from several disciplines to synthesize coherent solutions to the many challenges posed by HCAI?

This research topic stipulates that a foundation for HCAI needs to be at the level of a new underlying logical (reasoning) framework, in an analogous way that Classical Logic is the foundation or Calculus for Computer Science. Resting on the thesis that such a logical framework should be built on a solid understanding of human **cognitive reasoning**, and acknowledging the natural link of argumentation with human cognitive reasoning and human decision making at large, the present research topic explores the proposal of **Argumentation** as the foundation or Calculus for Human-Centric AI (Dietz et al.).

2 Call for papers: aim and scope

The aim of this call and its suggestion for the foundational role of argumentation in Human-Centric AI was to help bring together the wide variety of work on argumentation—ranging from argumentation in Philosophy and Ethics to the pragmatics of argumentative discourse in human debates—to understand how to synthesize a viable and robust basis for the development and use of HCAI systems. Systems that would meet their cognitive and ethical requirements, and integrate symbiotically, as expert or peer companions, within the human society, by complementing and enhancing the natural intelligence of humans.

3 Research Topic contributions

In addition to the paper that sets the scene for this Research Topic (Dietz et al.), another eight papers were accepted, ranging from results in theoretical work, presentation of own frameworks and setting the context of their work in relation to human-machine interaction in general or with respect to expert domains. Several of the papers have developed own

empirical studies serving as an evaluation metric for their frameworks (Albini et al., Kilic et al., Straßer and Michajlova).

Two distinct research directions can be identified among the contributing papers: a direction focusing on theoretical frameworks and development of own empirical studies (Albini et al., Bringsjord et al., Cramer and van der Torre, Straßer and Michajlova), and a direction focusing on the aspects of human-machine interaction and applications to expert domains (such as the medical domain or law) (Bikakis et al., Castagna et al., Kilic et al., Rotolo and Sartor). Yet, all contributions have in common that they agree on the importance of argumentation as foundations for human-centric AI.

3.1 Theoretical frameworks and development of own empirical studies

Albini et al. discuss properties of explanations in the context of descriptive accuracy. This implies that explanation contents need to be in correspondence with the internal working of the explained system. The authors provide a formal definitions of naive, structural and dialectical descriptive accuracy using the family of probabilistic classifiers as the context of their analysis. They evaluate their notions by several explanations methods and conduct studies with a varied selection of concrete probabilistic classifiers. Finally, the authors demonstrate how descriptive accuracy could be a critical component in achieving trustworthy and fair systems.

Bringsjord et al. present a new cognitive calculus, in which the central aspect concerns arguments that compete non-monotonically through time. Their framework captures well the three use-case studies, the Monty Hall problem, PERI.2 and the cognitive architecture ARCADIA. Finally, the authors specify seven desiderata for their framework.

Cramer and van der Torre introduce the naive-based argumentation semantics SCF2 and prove that it satisfies two new principles, which are not simultaneously satisfied by any argumentation semantics in the literature. Motivated by findings from empirical studies, these principles seem to correspond well to what humans consider a rational judgment on the acceptability of arguments.

Straßer and Michajlova present a framework for reasoning with higher-order uncertainty. This system integrates with deductive argumentation and can be adjusted to perform well under the so-called rationality postulates of formal argumentation. The authors provide several notions of argument strength, studied both meta-theoretically and empirically by discussing an own empirical study on evaluating argument strength in the context of higher-order uncertainty.

3.2 Human-machine interaction and application to expert domains

Bikakis et al. present a visionary paper on the problem of opinion overload in which they argue that it is possibly solvable by exploiting the structure of realistic arguments and understanding an arguer's intentions. The authors identify the main challenges and technological directions, ranging from understanding and formalizing realistic arguments and debates, and developing

appropriate models and methods to augmenting Web technologies with the ability to automatically process online arguments. They propose that the realization of this vision will revolutionize Web experience.

Castagna et al. develop EQR (Explanation-Question-Response) argument schemes to generate explanations for treatment advice given to patients in the medical domain using the chatbot, EQRbot. No machine learning algorithm is used, but EQRbot depends on a dynamic knowledge base which is constantly updated with the patient's data.

Kilic et al. focus on expectations and perceptions regarding the role of interaction behavior of a digital companion (with experts and non-experts) in the health domain. They present an empirical requirement elicitation study for an argumentation-based digital companion to support behavior change. The results show that the extent to which a digital companion challenges or supports a user's attitude argumentatively (based on argumentation schemes) can influence the user acceptance and the interaction itself.

Rotolo and Sartor show how explainable AI and legal theory can be modeled in an argumentation framework with structured arguments. The authors review literature of formal models of legal argumentation and investigate the formal connection between argumentation and explanation in law. Their core contribution is the clarification of the structure in normative reasoning of the concepts of justification and explanation through formal argumentation. They argue that the distinction between justification and explanation is pragmatical rather than structural.

Author contributions

ED: Writing—original draft, Writing—review & editing. AK: Writing—original draft, Writing—review & editing. LM: Writing—original draft, Writing—review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

ED was employed by Airbus Central R&T. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



OPEN ACCESS

EDITED BY

Fabrizio Riguzzi,
University of Ferrara, Italy

REVIEWED BY

Marco Lippi,
University of Modena and Reggio
Emilia, Italy
Francesco Santini,
University of Perugia, Italy

*CORRESPONDENCE

Emmanuelle Dietz
emmanuelle.dietz@airbus.com
Antonis Kakas
antonis@ucy.ac.cy
Loizos Michael
loizos@ouc.ac.cy

SPECIALTY SECTION

This article was submitted to
Machine Learning and Artificial
Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

RECEIVED 28 May 2022

ACCEPTED 16 September 2022

PUBLISHED 21 October 2022

CITATION

Dietz E, Kakas A and Michael L (2022)
Argumentation: A calculus for
Human-Centric AI.
Front. Artif. Intell. 5:955579.
doi: 10.3389/frai.2022.955579

COPYRIGHT

© 2022 Dietz, Kakas and Michael. This
is an open-access article distributed
under the terms of the [Creative
Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction
in other forums is permitted, provided
the original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution
or reproduction is permitted which
does not comply with these terms.

Argumentation: A calculus for Human-Centric AI

Emmanuelle Dietz^{1*}, Antonis Kakas^{2*} and Loizos Michael^{3,4*}

¹Airbus Central R&T, Hamburg, Germany, ²Department of Computer Science, University of Cyprus, Nicosia, Cyprus, ³Open University of Cyprus, Latsia, Cyprus, ⁴CYENS Center of Excellence, Nicosia, Cyprus

This paper aims to expose and analyze the potential foundational role of Argumentation for Human-Centric AI, and to present the main challenges for this foundational role to be realized in a way that will fit well with the wider requirements and challenges of Human-Centric AI. The central idea set forward is that by endowing machines with the ability to argue with forms of machine argumentation that are cognitively compatible with those of human argumentation, we will be able to support a naturally effective, enhancing and ethical human-machine cooperation and “social” integration.

KEYWORDS

argumentation, position paper, human-centric approach, Artificial Intelligence, formal foundations, learning, reasoning, cognition

1. Introduction

AI started as a synthesis of the study of human intelligence in Cognitive Science together with methods and theory from Computer Science.¹ The general aim was to formulate computational models of human intelligence, and implement systems based on these models to emulate the natural form of intelligence. This original motivation was placed on the side lines in most of the middle years (1980–2010) of AI, with the emphasis shifting to super-intelligent AI (Bostrom, 2014) that could go beyond the ordinary human problem-solving capabilities within specific application domains, such as large-scale Planning (Bonet and Geffner, 2001), Data Analysis, and Data Mining (Nisbet et al., 2018).

The last decade has witnessed a return to the early AI goal of understanding and building human-like intelligent systems that operate in a cognitively-compatible and synergistic way with humans.² This is largely driven by a growing market demand for AI systems that act as (expert) companions or peers of their human users. The reemergence of “old AI,” now called **Human-Centric AI (HCAI)**, aims to deliver services within the realm of natural or commonsense intelligence to support and enhance the users’ natural capabilities in tasks ranging from organizing their daily routine, to ensuring compliance

1 The Dartmouth workshop (<http://raysolomonoff.com/dartmouth/>), where the term Artificial Intelligence was introduced, was a joint meeting between scientists from the forming disciplines of Computer Science with Cognitive Science and other related areas.

2 The recent book by Lieto (2021) describes the evolution of AI from the perspective of its link to human cognition, from its birth to today’s developments.

TABLE 1 Major characteristics of HCAI systems.

HCAI Characteristics	Description
Human in the loop	At the level of design, development, and deployment of systems
Human-friendly behavior	Within the sphere of human-like modalities of interaction
Cognitive compatibility	At the different levels of its various groups of human users
Synergistic accountability	Explainable, contestable, and debatable operation and behavior
Embodiment of systems	In the physical, mental, and emotional human environment
Body-mind-like model	Of operation to sense, recognize, think, and act
Developmental nature	Of systems through a continuous process of learning and adapting from experience
Social integration	Transparently within the human society

with legal or policy requirements, or to acquiring a first self-appreciation of a potentially troublesome medical condition.³

This ambitious vision for HCAI sets a challenging list of desiderata on the high-level characteristics that HCAI systems should exhibit. Table 1 gives an overview of a list of these characteristics.

But perhaps the most important desired characteristic of HCAI systems, overseeing all others, is: Adherence to human moral values promoting the responsible use of AI.

These vital characteristics for the development of HCAI systems attest to the need for a **multi-disciplinary** approach that would bring together elements from different areas, such as Linguistics, Cognitive Psychology, Social Science, and Philosophy of Ethics, and would integrate those into viable computational models and systems that realize a natural human-like continuous cycle of interacting with an open, dynamic, complex, and possibly “hostile” environment, and naturally enhance and improve their performance through their experience of operation and their evolving symbiotic relationship with their human users.

Building such HCAI systems necessitates a foundational shift in the problem-solving paradigm that moves away from the strictness and absolute guarantees of optimal solutions that are typically adopted for conventional computing, which are often brittle and break down completely when new information is acquired. Instead, HCAI would benefit by adopting **satisficing solutions** that strike an acceptable balance between a variety

of criteria, are tolerant to uncertainty and the presence of incompatible alternatives, are robust across a wide range of problem cases, and are elastic in being gracefully adapted when they are found to have become inappropriate or erroneous in the face of new information.

This realization that intelligent solutions require the flexibility of accepting the possibility that errors can occur has been stated by Alan Turing, a forefather of Artificial Intelligence, at his lecture to the London Mathematical Society on 20th of February 1947 (Turing, 1947):

“[...] if a machine is expected to be infallible, it cannot also be intelligent.”

Accepting this realism of sub-optimal performance, HCAI systems would then use problem instances where they have experienced the fallibility of their current solutions to gradually adapt and improve the satisficing nature of those solutions.

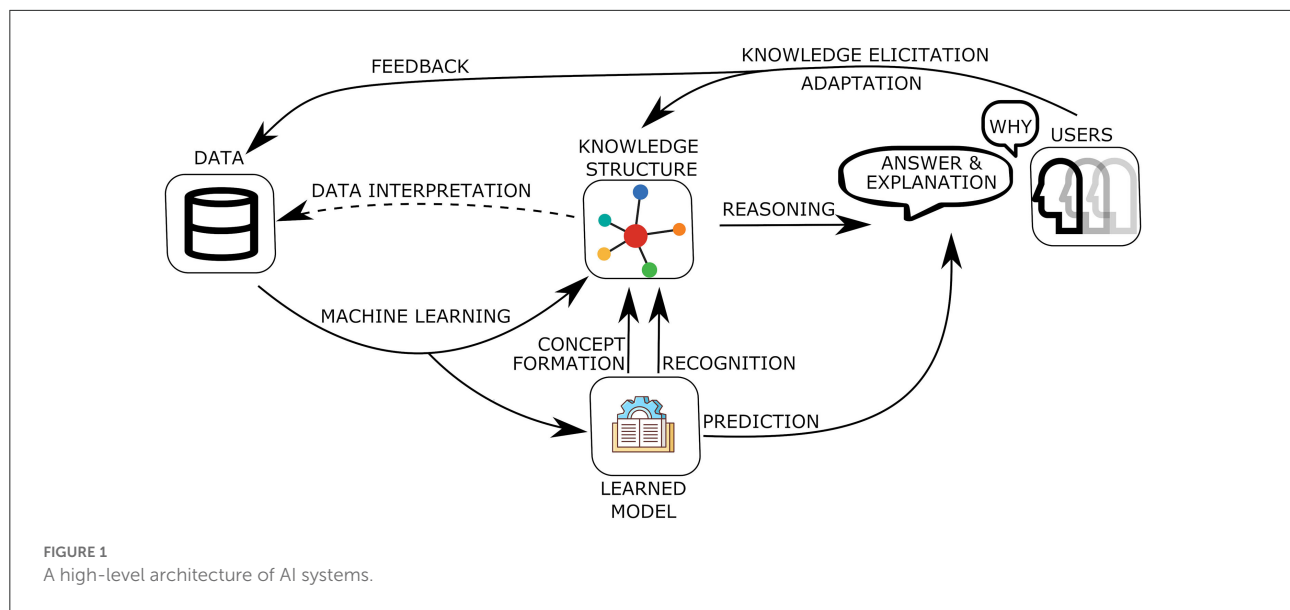
The nature of HCAI systems under a new paradigm of accepting and tolerating reasonably-good solutions suggests new perspectives on the Learning and Reasoning processes, which operate together in synergy to produce intelligent behavior: a **new reasoning** perspective as a method of analyzing the acceptability of possible alternative solutions; a **new learning** perspective as a process of generating knowledge that can resolve the ambiguity in the data, rather than knowledge that draws definite predictions or defines concepts.

Although we have described these as new perspectives, they have essentially been present in AI for some time. The new reasoning perspective of not always arriving at conclusive or best conclusions is implicitly assumed by the areas of Non-Monotonic Reasoning and Belief Revision, proposed from the very start of AI, as essential elements of reasoning that would need to differ from formal classical reasoning. Similarly, the new learning perspective underlies, for example, the Probably Approximately Correct (PAC) Learning theory, where it is explicitly recognized that one can typically only approximate what one learns.

The inability of the new forms of learning and reasoning to reach a definitive answer is compensated in HCAI systems by the provision of **explanations** of the satisficing alternatives, which offer an account of the lack of (or inability to reach) best answers. This explanation-based interaction needs to be **cognitively compatible** with the human users and developers of the systems, in order to facilitate the integration of the various processes and entities that exist within the application environment.

To help us place a human-centric perspective in today's terrain of AI research let us consider a typical high-level architecture of AI systems as shown in Figure 1. In this, learning and reasoning are tightly interconnected and both have a central role within the architecture. Learning is a continuous process that occurs throughout the life and operation of the system. **Machine learning** is used, e.g., in Deep

³ Today there are several centers dedicated to HCAI, such as <https://www.humane-ai.eu/>, <https://hai.stanford.edu/>, <https://human-centered.ai/>, <https://humaine.info>.



Neural Learning, to generate structures for direct prediction, typically lower-level akin to *system 1* (Kahneman, 2011) in human reasoning. This could be identifying or **recognizing** some property of the current state of the environment to be combined with the general knowledge of the system or indeed to output a predictive classification directly to the user. Machine learning is also used at the symbolic level to learn the structure of and populate the knowledge of the system that is to be used for higher-level, akin to *system 2*, cognitive reasoning by the system. Recently, there is a strong interest in the integration of sub-symbolic and symbolic learning so that through such methods we have an emergent **concept formation** process of identifying and forming high-level cognitive concepts on top of the sub-symbolic learned structures.

Together with learning we can also use methods of **knowledge elicitation** from experts and/or users of the system to build the knowledge of a system and the general structure that we want the knowledge to have for our system. This is particularly useful at the initial stages of the development of a system and helps us to steer the development along a general form that we desire. For example, knowledge elicitation can be used to provide the basic guidelines for moral and ethical behavior of the system, which could then be continuously refined and **adapted** during its operation from its experience of interaction with the outside environment of users and the society in which it operates.

The system's interaction with its environment, which includes its developers and users, goes beyond simply providing the answers of its reasoning or prediction. It engages into a dialogue based on **explanations** of the system's **answer** at a level compatible with the way the human **users** (to which the

explanations are addressed) themselves reason about the task. In order to have such meaningful interfaces, the **knowledge structure** of many AI systems is often connected to some structured form of Natural Language, so that its processing by the system can be linked to the human interpretation of the associated natural language form.

The development of an AI system is continuous with the **feedback** from its environment providing information to either revise and **adapt** the current state of its knowledge or to generate new data for further learning. For this development to be smooth it needs to take place under the prism of the current knowledge of the system. Hence, the results of reasoning by the system need to be explainable in terms of the current knowledge so that useful and meaningful feedback can be given to the system by its environment. Similarly, new experiences, that would drive new possibilities of learning, can first be interpreted under the current knowledge of the system to form suitable new data for further learning so that the new knowledge produced can fit naturally within the existing knowledge structure to which the system is committed. The development of the central element of the knowledge of an AI system is thus a matter of smooth evolution rather than a sequence of disconnected learning and adaptation processes.

Paper position: What is then an appropriate foundation for building HCAI systems with this variety of behavior characteristics and design features; a foundation that would give unity to the field and allow it to draw elements from several disciplines in order to synthesize coherent solutions to the challenge posed by HCAI?

We propose that such a foundation needs to be at the level of a new underlying logical framework, in an analogous way that Classical Logic is the foundation or Calculus for

Computer Science (Halpern et al., 2001). Resting on the thesis (or hypothesis) that this logical framework should be built on a solid understanding of human **cognitive reasoning**, and acknowledging the natural link of argumentation with human cognitive reasoning and human decision making at large, this paper proposes **Argumentation** as the foundation or Calculus for Human-Centric AI.

The **aim of this paper** and its suggestion for the foundational role of argumentation in Human-Centric AI is to help bring together the wide variety of work on argumentation — ranging from argumentation in Philosophy and Ethics to the pragmatics of argumentative discourse in human debates — in order to understand how to synthesize a viable and robust basis for the development and use of HCAI systems. Systems that would meet their cognitive and ethical requirements, and integrate symbiotically, as expert or peer companions, within the human society, by complementing and enhancing the natural intelligence of humans.

The rest of the paper is organized as follows. Section 2 presents the general features of argumentation in support of the position of the paper. Section 3 gives a brief overview of the main components of computational argumentation, formalization and pragmatics, and illustrates the role of argumentation in HCAI systems with two example systems. Section 4 analyzes the main challenges that would need to be faced by any logical foundation of HCAI, linking to these challenges the features of argumentation that would be relevant in addressing them. Finally, Section 5 concludes by briefly discussing the importance of an interdisciplinary approach to HCAI.

2. Why argumentation as a logical foundation?

We ground the proposal for argumentation as a suitable logical foundation of HCAI on two observations about argumentation and their connection with the historical development of Cognitive Science and Logic-based AI.

The first such observation is the strong cognitive support for argumentation and its link to different cases of human thinking. This stems from many studies in Cognitive Science and Psychology, and based on experiments and theories that have widely compared human informal reasoning with classical formal reasoning (Evans, 2010). The early motivation of these works was to examine how rational, i.e., how close to strict logic, human reasoning is, and to record its deviation from the valid formal logical reasoning. In recent years, the paradigm changed from such normative theories, of how humans “ought to reason,” to descriptive theories, of how humans “actually reason.” Despite significant differences between the observed informal reasoning and the strictly

valid formal reasoning, most humans are convinced that their way of reasoning is correct. Diverging from valid formal reasoning is often necessary to make intelligent decisions in everyday life!

An analogous shift can also be observed in Economics, from assuming the human as being “homo economicus,” i.e., an idealized rational agent in Neoclassical Economics, to accepting the bounded rationality of humans in Behavioral Economics, where the interest is in understanding how and why humans make decisions (Kahneman and Tversky, 1979; Paglieri and Castelfranchi, 2010) rather than modeling optimal choices. Decisions taken by people at large have been observed to deviate from logically strict or rational reasoning, and rather follow a heavily biased form of reasoning. Given the limited memory resources and time constraints of humans, the use of “efficient reasoning shortcuts,” such as biases or heuristics, are not only reasonable but necessary.

There is now strong evidence in various studies from Cognitive Psychology, brought together in the work of Mercier and Sperber (2011), that humans arrive at conclusions and justify claims by using arguments. With repeated experimental studies, Mercier and Sperber came to the conclusion that humans engage in motivated thinking through argumentation in order to defend their positions. In other words, argumentation is the “means for human reasoning.” Within the dual-process theory of human reasoning (Kahneman, 2011), with a *system 1* fast and intuitive process and a *system 2* slow and reflective process, Mercier and Sperber argue that “all arguments must ultimately be grounded in intuitive judgments that given conclusions follow from given premises,” in contrast to the usual assumption that *system 2* is unbiased and rather normative.

While in Cognitive Psychology and Behavior Economics the link to argumentation is examined following the scientific method of observation and theory formation, within the Humanities and particularly in Philosophy, scholars have been equating human informal reasoning with argumentation for centuries now. The entry on Informal Logic in the Stanford Encyclopedia of Philosophy (<https://plato.stanford.edu/entries/logic-informal/>) states:

“Though contributions to informal logic include studies of specific kinds or aspects of reasoning, the overriding goal is a general account of argument which can be the basis of systems of informal logic that provide ways to evaluate arguments. Such systems may be applied to arguments as they occur in contexts of reflection, inquiry, social and political debate, the news media, blogs and editorials, the internet, advertising, corporate and institutional communication, social media, and interpersonal exchange. In the pursuit of its goals, informal logic addresses topics which include, to take only a few examples, the nature and definition of argument, criteria

for argument evaluation, argumentation schemes, [...] and the varying norms and rules that govern argumentative practices in different kinds of contexts.”

Clearly, from the point of view of Humanities and other disciplines, human informal reasoning is a matter of argumentation.

The second main observation in support of argumentation concerns its relation and comparison with Classical Logic. The alternative of retaining Classical Logic, which has served conventional computing well over the decades, as the logical foundation for HCAI fails to capture fully certain forms of human reasoning that are well-outside the realm of formal classical logic. From the very early days of AI, the goal to address this discrepancy resulted in the search for and development of new logics for AI, such as non-monotonic logics, probabilistic, or fuzzy logics. In particular, a plethora of **non-monotonic logics** (Reiter, 1980; Shoham, 1987; Marek and Truszczyński, 1991) were proposed as candidates for the logical foundations of commonsense reasoning, starting with the logic of Circumscription for formalizing the Situation Calculus, a system for commonsense reasoning about the effects of actions and the change they bring about (McCarthy, 1968). These new logics aimed to capture the non-monotonicity feature of human reasoning, recognizing that, in contrast to formal Classical Logic, inferences should be flexible to missing or ambiguous information, and tolerant to (apparently) contradictory information, and should be possibly abandoned in the face of new relevant information.

Nevertheless, these new logics remained bound to the same formal and strict underpinning of Classical Logic making it difficult to deliver on their promise of “AI systems with commonsense” and human-like natural intelligence. On the other hand, the study of argumentation in AI, which was grounded on work in Philosophy and Cognitive Science (Toumlin, 1958; Perelman and Olbrechts-Tyteca, 1969; Pollock, 1987), showed that it was possible to reformulate (and in some cases extend) most, if not all, such non-monotonic AI logical frameworks (Bondarenko et al., 1997). Furthermore, it was recently shown that, within this AI approach to **Computational Argumentation**, it is possible to reformulate even Classical Logic reasoning as a special boundary case of argumentation, hence presenting argumentation as a universal form of informal and formal reasoning (Kakas et al., 2018; Kakas, 2019). These results together with the many links that Computational Argumentation has formed, over the last decades, with studies of argumentation in several other disciplines (see e.g., the journal of Argument and Computation⁴), have given a maturity to the field of Argumentation that allows it to serve as a candidate for the logical foundations of Human-Centric AI.

4 <https://www.iospress.com/catalog/journals/argument-computation>

3. Computational argumentation: An overview

In this section we present a brief overview of (Computational) Argumentation, highlighting its elements that are most relevant to its possible foundational role for Human-Centric AI systems. This overview is built by considering elements drawn from the large corpus of work on Argumentation in AI over the last few decades.⁵ It concentrates on the essential elements of argumentation as a general logical system of human cognitive reasoning (or thought), avoiding technical details that may vary over different approaches and that are not crucial for understanding the central link of argumentation and reasoning.

Argumentation is a process of debating the alternative positions that we can take on some matter, with the aim to justify or refute a certain standpoint (or claim) on the matter. It can take place socially within a group of entities, with each entity typically taking a different standpoint and arguing its case, or within a single entity that contemplates internally the various standpoints in order to decide on its own stance. The process is **dialectic**, where in the social context it is carried out *via* an **argumentative discourse** within Natural Language in a debate between the different entities, whereas in the individual case this is done within an introspective internal debate within the thinking entity.

The dialectic process of argumentation takes place by (i) starting with some argument(s) directly supporting the desired standpoint, then (ii) considering the various counter-arguments against the initial argument(s), and (iii) defending against these counter-arguments, typically with the help of other arguments as allies of the initial arguments. The process repeats by considering further counter-arguments against these new allied defending arguments. We therefore have an “argumentation arena,” where arguments attack and defend against each other in order to support their claims, and the aim is to form a **coalition (or case) of arguments** that collectively supports “well” a desired standpoint. In forming such a coalition, we may need to include arguments that do not refer directly to the primary matter in question, but refer to secondary matters that have come into play through the initial stages of the argumentation process.

This arena of argumentation can be captured by a formal **argumentation framework**, which in an abstract form is a triple $\langle \text{Args}, \text{Att}, \text{Def} \rangle$, where *Args* is a set of arguments, *Att* is an **attack (or counter-argument)** binary relation between arguments, and *Def* a **defense (or defeat)** binary relation

5 Work in the area of Computational Argumentation can be found in the journal of Argument and Computation and the International Conference on Computational Models of Argument (COMMA). Other sources for review material of the area include (Bench-Capon and Dunne, 2007; Simari and Rahwan, 2009; Atkinson et al., 2017; Vassiliades et al., 2021).

between arguments. Typically, the defense relation Def is a subset of the attack relation Att capturing some notion of the relative strength between the attacking arguments. Hence when $(a_1, a_2) \in Def$ the argument a_1 is strong enough to defend against (or defeat) a_2 .

In practice, abstract frameworks are realized by structured argumentation frameworks (Kakas and Moraitis, 2003; Gracia and Simari, 2004; Prakken, 2010; Modgil and Prakken, 2013), expressed as triples of the form $\langle \mathcal{A}s, \mathcal{C}, \succ \rangle$, where $\mathcal{A}s$ is a set of (parameterized) **argument schemes** (Walton, 1996), instances of which form the arguments, \mathcal{C} is a **conflict relation** between argument schemes (and between their arguments), and \succ is a **priority (or preference or strength) relation** between argument schemes (and between their arguments). A structured argumentation framework, $\langle \mathcal{A}s, \mathcal{C}, \succ \rangle$ forms a **knowledge representation** framework, where knowledge is represented in a structured form, and on which the dialectic argumentation process of attack and defense can be performed.

Argument schemes⁶ in $\mathcal{A}s$ are parameterized named statements of association between different pieces of information. They can be represented in the simple form of $\mathcal{A}s = (\text{Premises} \triangleright \text{Position})$, associating the information in the Premises with the statement of the Position. Hence, given the information in the Premises we can construct an argument (or reason) supporting the Position (or Claim) based on the link from the Premises to the Position in the argument scheme. The attack relation between arguments is constructed directly from the conflict relation \mathcal{C} , which normally stems from some expression of incompatibility, e.g., through negation, in the underlying language of discourse. The defense relation is built using the priority relation \succ , where, informally, an argument defends against another argument if and only if they are in conflict and the defending argument is not of lower priority than the argument it is defending against. Importantly, and in contrast to the conflict relation which is static, the priority relation is **context-sensitive**, and depends crucially on (how we perceive) the current state of the application environment.

In computational argumentation, we impose a **normative** condition on which argument coalitions are considered **acceptable** as a **valid case** of support for their corresponding standpoints. This normative condition of acceptability stems directly from the dialectic argumentation process to examine and produce cases of support. Informally, an **acceptable**

argument coalition is one that can defend against all its counter-arguments while not containing an internal attack between (some of) the arguments within the coalition⁷. In other words, attacking (or counter) arguments should be defended against, but in doing so we cannot introduce an internal attack between the arguments of the coalition.

This normative condition of acceptability of arguments gives a logical structure to argumentation. In comparison with Classical Logic, the **Logic of Argumentation** replaces the underlying structure of a truth model with that of an acceptably valid case of arguments. Logical conclusions are drawn in terms of the valid cases of arguments that support a conclusion. When a valid case supporting a conclusion exists we say that this is a **plausible or possible conclusion**. If, in addition, there are no valid cases for any contrary conclusion, then we have a **definite conclusion**.

Clearly, definite conclusions are closer to logical conclusions of formal logical reasoning systems, like that of Classical Logic. When they exist, definite conclusions are based on clear winning arguments in the argumentation arena, which ensure the strict and absolute consequence of the conclusion. This, then, corresponds to the **strict rationality** form of formal logical reasoning. For example, in the context of a decision problem where we require from the logic to identify rational choices for our decision, these definite conclusions would correspond to optimal choices. The Logic of Argumentation allows, in addition, a softer form of **Dialectic Rationality**, where several, typically opposing, conclusions (e.g., decisions) are considered rational as they are **reasonably justified** by an argument case that is valid. We thus have a more general form of rationality where the absolute guarantees of classical strict rationality are replaced by the accountability of dialectic rationality *via* the provision of a **justification** for the conclusion or choice. These justifications contain, in a transparent and explicit way, the different arguments that would render a conclusion **reasonable**.

Dialectic rationality depends on the **relative** importance we place on the various requirements of the problem at hand and the relative “subjective” value we give to the relevant information. Thus, a decision can be accepted as rational when it is reasonable under some set of standards or requirements, including the subjective preferences or biases that we might have for a specific standpoint. Concerns about a specific choice and the beliefs that underlie this are addressed in the dialectic argumentation process that has produced the argument coalition supporting that choice. Importantly, if new concerns are raised, e.g., by the dynamic application environment, then these should be addressed, and if the argument coalition for the choice cannot be adapted to address these concerns, i.e., to defend against the counter-arguments they raise, then the

⁶ Argument schemes are central to argumentation. They have been extensively studied, starting with Aristotle in his books of Topics, in various contexts of human argumentation. In recent times they are several important works that aim to standardize their form (Toumlin, 1958; Walton, 1996; Walton et al., 2008). In the work of Wagemans (2018) a periodic table for classifying the different arguments used by people is proposed.

⁷ More generally, an acceptable argument coalition is one that once adopted can render all its counter-arguments non-acceptable.

rationality of the choice is lost and as a consequence the suitability of the solution is lost.

3.1. Pragmatic considerations of argumentation

The feature of the Logic of Argumentation to naturally provide a justification for its conclusions is very useful within the **social context** of application of systems, as the justification can be turned into, and presented as, an **explanation** for the conclusion. The issue of providing explanations for the results of AI systems is today considered to be a major requirement for any AI system, and forms the main subject matter of **Explainable AI**. Explanations of conclusions, or taken decisions, serve well their social role of interaction when they give the basic reasons of support (attributive), they explain why a conclusion is supported in contrast to other opposing conclusions (contrastive), and they provide information that guides on how to act following the conclusion (actionable) (Miller, 2019).

Argumentation is naturally linked to explanation the recent surveys of Čyras et al. (2021) and Vassiliades et al. (2021) as well as the proceedings of the recent, first, International Workshop on Argumentation for Explainable AI (ArgXAI)⁸ give a thorough exposition of this link and its potential significance in AI. The arguments justifying a decision can form the basis of an explanation to another party. The argumentative dialectic reasoning process and the acceptable coalition of arguments that it constructs can be unraveled to give an explanation. Such explanations extracted from an acceptable argument coalition have an **attributive** element coming from the initial arguments that support the conclusion, while the defending arguments against the counter-arguments will provide a **contrastive** element of the explanation. These arguments also point toward taking (further) actions to confirm or question their premises, particularly when these relate to subjective beliefs or hypotheses.

As described above, the theoretical notion of **computation** that stems from the Logic of Argumentation, is that of the (iterative) dialectic argumentation process of considering arguments for and against an initial conclusion and other subsidiary conclusions that help to defend the arguments supporting the initial conclusion. During this dialectic process we have (at least) three choices that can render the process computationally intensive and highly complex. These complexity points are: the choice of initial argument(s), the choice of counter-arguments, and finally the choice of the defending arguments. The consideration of the **pragmatics of argumentation** (van Eemeren and Grootendorst, 2004) thus becomes an important issue when argumentation is applied in the real world. This includes questions of how are arguments

activated and brought to the foreground of the argumentative process, and similarly how is the relative strength of arguments affected by the changing state of the external environment in which the process takes place.

To address this issue of the pragmatics of argumentation, we can draw from the large body of work on **Human Argumentation**, which studies how humans argue and how this results in the effectiveness that we observe in human reasoning. This study starts from Aristotle in the books of *Topics*, where he attempts to systemize argumentation and give detailed prescriptions of good practices for the way one can argue for or against a position. Recently, over the past decades, several works have set out detailed methods for formulating and understanding human argumentation from various different perspectives: philosophical, linguistic, cognitive, and computational; see the work of van Eemeren et al. (2014) for a comprehensive review. These include studies of understanding the various types of argument schemes that humans use in their argumentative discourse (Toulmin, 1958; Walton, 1996; Walton et al., 2008), or how the process of human argumentation relates to human reasoning (Pollock, 1987), and how human argumentation discourse can be regulated by pragmatic considerations that can help lead to agreement or a resolution of different standpoints in a debate (van Eemeren and Grootendorst, 2004).

Cognitive principles can then be drawn from these studies and from the study of human reasoning more generally, to be used as “cognitive guidelines” within the formal computational frameworks of argumentation to give a form of **Cognitive Machine Argumentation** that would be cognitively compatible with the argumentation and reasoning of humans (Saldanha and Kakas, 2019; Dietz and Kakas, 2021). This can then support an effective human-machine interaction *via* compatible forms of argumentation between machine systems and their human users.

Human argumentation is typically carried out in a social setting, as an argumentative discourse in Natural Language. It is, therefore, important to be able to recognize and extract the argumentation structure from the natural language discourse (Hinton, 2019, 2021). This includes the ability to recognize which parts of text are indeed argumentative, to identify the quality of the arguments that are extracted from the text, and, more generally, to extract the argumentative structure of support and attack between arguments extracted from various parts of some piece of text under consideration.

Argument mining is an area of study of argumentation which has strong links both with computational argumentation and with the study of human argumentation. It aims to automate the process of extracting argumentative structure (Lippi and Torroni, 2016; Lawrence and Reed, 2019) from natural language. It combines elements from the various different studies of human argumentation with methods from computational linguistics in order to turn unstructured text into

⁸ <https://people.cs.umu.se/tkampik/argxai/2022.html>

structured argument data. This is typically carried out using an ontology of concepts relevant to some specific area of (human) argumentative discourse that we are interested in. Then applying argument mining on corpora of textual information related to a particular problem domain forms an important method to populate a computational argumentation framework for a corresponding application domain of interest.

Having described the basic idea behind Computational Argumentation and certain important connections to relevant lines of work, let us now illustrate, through two examples of candidate AI systems, how the Logic of Argumentation connects with Human-Centric AI. How would the Logic of Argumentation provide the basis for formulating and solving a Human-Centric AI problem?

3.2. Everyday assistants: Cognitive consultation support

Let us first consider the class of **Cognitive Review Consultation Assistants**, and more specifically a **Restaurant Review Assistant**, whose main requirement is to help human users to take into account the online reviews available on the various options in some decision problem. For simplicity, we will concentrate on how the logic of argumentation can help us use the information in the reviews for one particular restaurant in order to form a personal opinion about this restaurant. The problem of the assistant is to evaluate, but not necessarily to decide, whether the restaurant in question is a **reasonable choice** or not for a personal user of the system. A solution is an informed explanation of why the restaurant is a reasonable choice or not for the user based on the information on the reviews. Furthermore, we are not interested in identifying if a restaurant is an optimal best choice for us to dine out but rather a satisficing choice.

How can we represent this problem of the Restaurant Review Assistant in terms of an argumentation framework $\langle \mathcal{A}, \mathcal{C}, \succ \rangle$? The argument schemes or arguments for and against a restaurant can be built using as premises the different types of information that the reviews contain. We will consider a **simple form of argument schemes** where these consist of a named association between a set of premises and an atomic statement of the supported position. To start with, the overall score of the reviews provides the premise for the basic arguments for the deliberation of the assistant: if the overall score is above some (personal) high threshold this will form an argument in favor of the restaurant, and if it is below some (personal) low threshold this will form an argument against the restaurant:

$$\begin{aligned}\mathcal{A}s_1 &= (\text{HighScore} \triangleright \text{Favorable}) \\ \mathcal{A}s_2 &= (\text{LowScore} \triangleright \text{Non_Favorable}).\end{aligned}$$

HighScore means that the score is above the high threshold, and *LowScore* that it is below the low threshold. Furthermore, when the overall score is in between these thresholds then we can have another two basic arguments, one supporting the position *Favorable*, and the other supporting *Non_Favorable*:

$$\begin{aligned}\mathcal{A}s_3 &= (\text{MiddleScore} \triangleright \text{Favorable}) \\ \mathcal{A}s_4 &= (\text{MiddleScore} \triangleright \text{Non_Favorable}).\end{aligned}$$

To complete the representation of the problem, we include in the conflict relation the obvious conflict between arguments that support the incompatible positions *Favorable* and *Non_Favorable*, and we leave the priority relation between these four arguments empty. In fact, the mutual exclusivity of the premises between most of the pairs of arguments, except between $\mathcal{A}s_3$ and $\mathcal{A}s_4$, makes the need to consider possible relative priorities essentially unnecessary. For the pair of $\mathcal{A}s_3$ and $\mathcal{A}s_4$, it is natural not to assign a relative priority between them. Hence, all conflicting arguments attack and defend against each other.

In general, the reviews will refer to, and comment positively or negatively on, properties that we usually consider relevant in evaluating the suitability of a restaurant: “service,” “cost,” “quality or quantity of food,” “atmosphere,” etc. Each such review would thus generate arguments for and against the suitability of the restaurant according to argument schemes of the following general form:

$$\begin{aligned}\mathcal{A}s_{+ve}(\text{Review}(\text{Id})) &= (\text{Positive}(\text{Property}) \triangleright \text{Favorable}) \\ \mathcal{A}s_{-ve}(\text{Review}(\text{Id})) &= (\text{Negative}(\text{Property}) \triangleright \text{Non_Favorable}).\end{aligned}$$

The premises of the resulting arguments are the positive or negative opinions that a review expresses on some of these relevant properties.

In general, the priority relation between these arguments would be mostly affected by the personal preferences of the human user, as communicated to their customized personal assistant, possibly through Natural Language guidelines, such as: *I prefer to avoid expensive restaurants, but I like to eat quality food.* With this statement, the user has identified the properties of “cost” and “quality” of food to be of particular relevance and importance, giving corresponding priority to arguments that are built with premises referring to these properties. Hence, a review that considers the restaurant expensive will give an argument built from $\mathcal{A}s_{-ve}(\text{Review}(\text{Id}))$ higher priority than (some of the) other arguments for the position *Favorable*. But, as the guideline indicates, this argument will not have higher priority than arguments built using the scheme $\mathcal{A}s_{+ve}(\text{Review}(\text{Id}))$ from reviews that stress the high quality of the food.

Given the aforementioned arguments, the dialectic argumentative reasoning simulates a debate between the various reviews (or possibly only a subset of the reviews chosen according to some criteria) and their positive and negative comments. Regardless of whether the assistant reaches a definite

conclusion or remains with a dilemma on being favorable or not toward a given restaurant, the assistant will be able to provide an explanation based on the supporting arguments and the dialectic debate that has resulted in the acceptability of the argument according to the wishes of the user. These explanations will be very useful in the process of the assistant gaining the trust from its human user.

Cognitive Review Consultation Assistants are quite focused on very specific topics of interest. At a more varied level, we may want to build HCAI systems of “Search Assistants” to help us in getting a reliably balanced understanding on a matter that we are interested in. Eventually, Search Assistants should extract the arguments for and against the matter that we are interested in, together with their relative priorities, presenting to us a balanced view of the dialectic debate between these arguments. Tools and techniques from argument mining are directly applicable on, and a natural fit for, this extraction task, as one seeks to understand the argumentative discourse expressed in Natural Language, be that in the statements made by the human user in communicating their search parameters and preferences, or in the text or reviews that are being searched. For example, in the Reviews Assistant case, argument mining can be used (Cocarascu and Toni, 2016) to extract from the text of the reviews the arguments they are expressing, as well as the relative strength between these arguments, in support of positive or negative statements on the various features that are relevant for the user who is consulting the system.

3.3. Expert companion: Medical diagnosis support

Let us now consider another example class of Human-Centric AI systems, that of **Medical Diagnosis Support Companions**. This class of problems differs from the previous example of Everyday Assistants in that these systems are based on expert knowledge, on which there is large, but not necessarily absolute, agreement by the expert scientific community. Furthermore, these systems are not personalized to individual users, but they can have different groups of intended users. Their general aim will then depend on their user group. For example, if the user group is that of junior doctors in some specialization who need to train and gain practical experience in their field, then, within the framework of Human-Centric AI, these systems can have the general overall aim to:

“Support clinicians feel more confident in making decisions, helping to avoid over-diagnosis of common diseases and to ensure emergency cases are not missed out.”

Medical diagnostic knowledge that associates diseases with their observable symptoms can be

represented in terms of argument schemes of the general form:

$$\mathcal{A}s = (\text{Symptoms} \triangleright \text{Disease}).$$

Hence, based on the premise that the information in *Symptoms* holds, we can build an argument that supports a certain disease (as the cause of the symptoms). For different sets of symptoms we would then have argument schemes that would provide arguments that support different diseases. These associations are expertly known and are treated as arguments, which means that they are not understood as definitional associations that must necessarily follow from the symptoms. Rather, for the same set of symptoms we can have argument schemes supporting different diseases, rendering each one of these diseases as plausible or suspicious under the same set of premises.

To complete the representation of the problem knowledge within an argumentation framework $\langle \mathcal{A}s, \mathcal{C}, \succ \rangle$, we would need to specify, in addition to these argument schemes, the conflict and priority relations. The conflict relation would simply capture the information of which diseases do not typically occur together. The priorities of arguments can come by following the diagnostic process followed by doctors in their practice of evidence-based medicine: Argument schemes as above apply on initial symptoms, e.g., the presenting complaints by a patient. Then the doctors have contextual knowledge of further symptoms or other types of patient information that allows them to narrow down the set of suspected diseases. This can be captured within the argumentation framework in terms of giving relative priority between the different basic argument schemes, where the priority is conditional on some extra contextual information.

In fact, one way to capture this contextual priority is in terms of preference or priority argument schemes, which support the preference of a basic argument for one disease over another basic argument for another disease, of the form:

$$\mathcal{A}s_{\text{prefer}} = (\text{Context} \triangleright (\mathcal{A}s_1 \succ \mathcal{A}s_2)),$$

where $\mathcal{A}s_1$ and $\mathcal{A}s_2$ are argument schemes supporting different diseases based on the same or overlapping premise information of symptoms and patient record.

Typically, the dialectic argumentation process would start between basic arguments supporting the alternative possible diseases, but then this is **entangled** with other dialectic argumentative processes arguing for the priorities of those basic arguments, and thus their ability to attack and defend, and so on. Hence, depending on the extra contextual information that is received by, or actively sought from, the environment, and the preference arguments that are enabled as a result, some of the diseases which were acceptably supported at the basic (general) level will not be so any more, if they are attacked by arguments supporting other diseases but with no defense available as before.

Therefore, the set of suspicious diseases will be reduced, and the overall result will be that the diagnosis is further focused by this extra contextual information.

Another type of knowledge that can focus the result of the diagnostic process is contra-indication information, which supports the exclusion of some specific diagnosis. Such contra-indication information is typically strong and overrides other contextual information that would render a specific disease as being suspicious. This can be captured within argumentation in a similar way as above, by argument schemes that give priority to arguments against a specific diagnosis.

It is natural to compare this argumentation-based approach to medical diagnosis support systems with that of medical expert systems (Buchanan and Shortliffe, 1984) that were popular in the early days of AI. The knowledge in those early systems had to be carefully crafted by the computer scientists in terms of strict logical rules. Those rules, like the argument schemes we have described above, linked the symptoms to diseases⁹. The difference, though, with the argumentation-based representation, is that expert systems try to represent the knowledge in terms of logical definitions of each disease, a task which is very difficult, if not impossible, exactly because of the contextual differences that such definitions must take into account. For example, as definitions those rules would need complete information, and would need to ensure that there is no internal conflict or inconsistency among them.

The argumentation-based representation, on the other hand, can be incrementally developed by modularly adding new expert knowledge or by taking into consideration the feedback. This more flexible approach to knowledge representation is linked to the different perspective of HCAI systems, away from the expert systems perspective of reproducing and perhaps replacing the human expert, and toward the perspective of keeping the “human in the loop,” where the systems aim to complement and strengthen the human expert’s capabilities.

4. Major challenges for Human-Centric AI

We now continue to describe some of the major challenges for the underlying logical foundations of Human-Centric AI and comment on how argumentation, in its role as a candidate for these foundations, relates to these challenges. We focus on presenting challenges at the underlying theoretical level of Human-Centric AI that would provide the basis for the principled development of systems, while we acknowledge

that many other, more particular, technological challenges, would also need to be addressed to achieve the goals of Human-Centric AI.

The challenges for Human-Centric AI are not new for AI, but they reappear in a new form adapted to the human-centric perspective of HCAI. Overall, the main challenge for HCAI, and for AI more generally, is to acquire an understanding of human intelligence that would guide us to form a solid and wide-ranging computational foundation for the field. In particular, we need to understand thoroughly **Human Cognition**, accepting that the process of cognition, and its embodiment in the environment, form the central elements of intelligence.

This understanding of human cognition includes the following three important aspects: (1) how cognitive knowledge is organized into concepts and associations between them at different levels, and how cognitive human reasoning occurs over this structured knowledge, (2) how cognitive knowledge is acquired and learned, and how the body of knowledge is improved or adapted through a gradual and continuous development process, and (3) how the internal integrated operation of cognition, from low-level perception to increasingly higher levels of cognition, is supported by an appropriate architecture, and how an individual’s cognition is integrated with the external physical and social environment. Below we will analyze separately these main challenge areas and discuss the inter-connections between them.

4.1. Knowledge and inference

Human-Centric AI systems are knowledge intensive. As in the case of human cognition, they will need to operate on large and complex forms of knowledge. To achieve this we need a framework for representing and organizing knowledge in structures that would facilitate appropriate types of inference and decision making. From one point of view (the anthropomorphic design and operation of AI systems), the task is to match the main features of Human Cognitive Knowledge and Reasoning, including their **context-sensitive** nature and the **multi-layered knowledge structure** into concepts and associations between them at different **levels of abstraction**.

The need for these characteristics of knowledge and reasoning had been identified from the early stages of AI, with various knowledge structures being proposed to capture them. For example, the structure of frames (Minsky, 1981) aimed to capture the context sensitive nature of knowledge. Similarly, inheritance networks (Horty et al., 1990) were used to capture the different cognitive levels of knowledge and a form of contextual inference based on hierarchical generalizations. Another such structure, that of scripts (Schank and Abelson, 1975), aimed to capture the context-sensitive nature of commonsense reasoning with the knowledge of stereotypical sequences of events, and the change over time

⁹ Note that this non-causal direction of association between symptoms and disease is the natural one when the knowledge is used in the practice of medicine, where doctors carry out the diagnostic process. The causal direction of association from a disease to symptoms is the natural direction when we are studying the underlying medical scientific theory.

that these events bring about. This approach of defining explicitly cognitive knowledge structures was replaced, over several decades up to the start of the 21st century, to a large degree by the search for **non-monotonic logics**. The emphasis was shifted away from suitable explicit structures in knowledge and the cognitive nature of the process of inference to that of rich semantics for these logics that would capture the intended forms of human cognitive reasoning. Intelligent reasoning would follow from the correctness of choice of the rich logical formalism.

Essentially, all these approaches were concerned with the major problem of the necessary adaptation of inference over different possible contexts. This challenge, named the **qualification problem**, was concerned with the question of how to achieve context-sensitive inference without the need for a complete explicit representation of the knowledge in all different contexts, and how this is linked to the desired inferences in each one of these explicitly represented general and specialized contexts. To address this problem of knowledge and reasoning qualification in non-monotonic logics, we would typically include some form of modalities and/or some semantic prescription in a suitable higher-order logic, typically over classical logic. The practical problem of turning the logical reasoning into a human-like cognitive inference in an embodied environment was considered to be of secondary difficulty by most of these approaches with some notable exceptions, e.g., in that of McDermott (1990).

Our proposal of argumentation as the logical calculus for Human-Centric AI assumes that an appropriate cognitive structure of knowledge can be captured within structured argumentation frameworks. This structure is given by the priority relation amongst the individual argument schemes, which expresses in the first place a direct and local form of qualified knowledge. This then induces implicitly a global structure on the knowledge *via* the attack and defense relations of argumentation that emerge from the locally expressed strength and conflict relations. The dialectic argumentative reasoning over this structure gives the qualification of inference over the various different and complex contexts. Indeed, Computational Argumentation, with its new approach to logical inference, was able to offer a unified perspective on these central problems of context-sensitive and qualified inference, by reformulating (and in many cases extending) most, if not all, known logical frameworks of non-monotonic reasoning in AI (Bondarenko et al., 1997).

The challenge for argumentation is to build on this, and understand more concretely the **argumentative structure of cognitive knowledge**, and how to use it to match the **practical efficacy** of human cognitive reasoning. For example, how do we recognize the context in which we are currently in so that we can debate among alternatives that are available in this context? Similarly, how do we recognize that there is insufficient current information that would lead to a reasonable inference? For

example, there might be too many different conclusions that are equally supported, and hence we seamlessly recognize that it is not worth examining the inference, and it is better to wait for further information. This is akin to what humans naturally do in understanding narratives, where we leave empty pieces in the picture or model of comprehension, waiting for the author to reveal further information.

Another challenge related to the cognitive structure of knowledge is the need for a natural link to **explanations** for the inferences drawn at different cognitive levels of abstraction. In the organization of knowledge we can distinguish concepts that typically need explanation and those which do not — a separation that is also context sensitive depending on the purpose of the explanation and on the audience receiving the explanation. For example, the recognition of an image as a case of some abstract concept, e.g., of Mild Cognitive Impairment, can be explained in terms of some lower level features of the image, e.g., small HIP volume, which normally do not require (or for which one does not normally ask for) explanation. Perhaps one could ask for an explanation of “small” and be given this by some numerical threshold, in which case the even lower level feature of being less than the threshold is unlikely to be further questioned for an explanation. In any case, explanations need to be cognitively compatible with the user or process to which they are addressed, i.e., expressed at the same level of understanding and within the same language of discourse.

Argumentation has a natural link to explanation. Premises of arguments directly provide an attributive element of an explanation, while the structure of the dialectic argumentative process can be used to form a contrastive part of the explanations, i.e., explain why some other inference or decision was not made. This link of argumentation to explanation and the general area of Explainable AI has recently attracted extensive attention by the computational argumentation community (Kakas and Michael, 2020; Ćyras et al., 2021; Vassiliades et al., 2021). The challenge is how to turn argumentation into the language of explanation in a way that the explanations are provided at an appropriate cognitive level and are of **high quality** from the psychological and social point of view, e.g., they are naturally informative and non-intrusively persuasive (Miller, 2019). Argumentative explanations can help the receiving process or human to take subsequent rationally-informed decisions, based on transparent attributive reasons for the rationality of a choice, while at the same time not excluding the freedom of considering or deciding on other decisions that are alerted to by the contrastive elements of explanations.

The high-level medium of human cognition, as well as the intelligent communication and interaction between humans, is that of **Natural Language**. The above challenges on the Structure and Organization of Knowledge and Reasoning need also to be related and linked with Natural Language as the medium of Cognition and Intelligence. Computational Linguistics and comprehension semantics and processes that

are context-sensitive, such as the distributed semantics of Natural Language, are important in this respect to guide the development of AI. At the foundational level, the challenge is to understand cognitive reasoning on the medium of Natural Language. How is the process of human inference grounded in Natural Language, as it is studied, for example, in Textual Entailment (Dagan et al., 2009)? Several argumentation-based approaches study this question by considering how argumentative knowledge (arguments and strength) are extracted or mined from natural language repositories (Lippi and Torroni, 2016; Lawrence and Reed, 2019), i.e., how argument schemes are formed out of text (Walton, 1996), or how we can recognize good quality arguments (Hinton, 2019, 2021) from their natural language expression. The foundational challenge for argumentation is to understand how, in practice, the process of dialectic argumentation relates to and can be realized in terms of a human-like argumentative discourse in Natural Language.

4.2. Developmental nature

The recognition of the central role that knowledge plays in Human-Centric AI systems comes with the challenge of how that knowledge comes about in the first place, and how it remains current and relevant across varying contexts, diverse users interacting with the systems, and shifting and dynamic circumstances in the environment within which the systems operate. And all these, while ensuring that the knowledge is in a suitably structured form to be human-centric. Depending on the eventual use of knowledge, different ways of acquiring that knowledge might be pertinent.

In terms of a first use of knowledge, Human-Centric AI systems need to have access to background knowledge, through which they reason to comprehend the current state of affairs, within which state they are asked to reach a decision. Such knowledge can be thought to be of a commonsensical nature, capturing regularities of the physical or social world. Trying to fit empirical observations into a learned structured theory would be akin to trying to cover a circle with a square. The language of learning needs to be flexible enough to accommodate for the fact that not all empirical observations can be perfectly explained by any given learned theory. As obvious as this might sound, the majority of modern machine learning approaches implicitly ignore this point, and rather proceed on the assumption that the learned theory is a total mapping from inputs to outputs. As a result, these learning approaches are forced to consider richer and richer representations for learned theories (e.g., in the form of deep neural networks with millions of learning parameters to tune) that can, in principle, fit perfectly the learned data, losing at the same time the structure that one would wish to have in the learned theories, and opting for optimal rather than satisficing

accuracy in their predictions at the expense of sub-par rather than satisficing efficiency.

An argumentation-based learned model, on the other hand, explicitly acknowledges that the learned theory only partially captures, in the form of sufficient conditions, whatever structure might be revealed in the empirical observations, choosing to abstain from predictions when these sufficient conditions are not met (e.g., for the areas of the circle that our outside the square). This is taken a step further, with these sufficient conditions not being interpreted strictly, but being defeasible in the presence of evidence to the contrary effect. Additional arguments in the learned model can thus override and fine-tune the conditions of other arguments (e.g., by pruning the corners of the square that might fall outside the circle).

By acknowledging the unavoidable incompleteness of a learned theory, a further related challenge emerges: the ability of a partially-good theory to be gracefully extended to a better one, without having to undertake a “brain surgery” on the existing theory. This elaboration tolerance (McCarthy, 1968) property allows one to adopt a developmental approach to learning, spreading the computationally demanding process of learning across time, while ensuring that each current version of the theory remains useful, usable, and easily improvable. An argumentation-based learned model can meet these requirements, as it can be gracefully extended with additional arguments, whose inclusion in the learned model is handled by the semantics of argumentation, without the need to affect the pre-existing theory. In case the extended part of the learned model comes in conflict with the original part, argumentation records that as a dilemma, and gives the learning process additional time to resolve this dilemma, even guiding the learning process on where it should focus its attention to be most effective.

In terms of a second use of knowledge, Human-Centric AI systems need to have access to decision-making knowledge, through which they reason to reach a decision on how to act in the current state of affairs, after comprehending that state with the aid of background knowledge. Such knowledge can be thought to be domain- and user-specific, capturing the preferences of the users of the system. It is expected, then, that such knowledge can be acquired by interacting with the users themselves whose preferences one wishes to identify.

In such an interaction, the system needs to employ a learning process that acknowledges the nature of human preferences, and the mental limitations of humans when communicating their preferences. Preferences might be expressed in a hierarchical manner (e.g., stating a general preference of red wine over white wine), with more specific preferences overriding the general preferences in certain contexts (e.g., when eating fish). Any preferences communicated by humans should, therefore, be taken as applicable in the absence of other evidence, but need to support their flexible overriding

in the presence of exceptional circumstances or specific contexts.

At the same time, the preferences expressed by a human undertaking the role of a coach for the learner (Michael, 2019) should support their juxtaposition against social norms, ethical principles, expert knowledge, and applicable laws. Irrespective of whether such norms, principles, and laws are learned or programmed into a Human-Centric AI system, it should be easy to integrate them with the user's preferences that are passively learned or more directly provided by the user to the learner.

Since humans communicate most often in natural language, either with the explicit aim of offering their knowledge to a specific individual, or as part of supporting their position against another in a dialectical setting (e.g., in a debate in an online forum), the process of knowledge acquisition should be able to account for natural language as a prevalent source of knowledge. Techniques from argument mining (Lippi and Torroni, 2016; Lawrence and Reed, 2019) can be used to extract arguments directly from human discourse expressed in natural language. This discourse could represent the dialogue that a human has with the machine, in the former's effort to communicate their preferences to the latter. Equally importantly, the discourse can be undertaken in a social context among multiple humans. Mining arguments from such a discourse could help identify arguments in support and against diverging opinions on a matter, commonly agreed upon norms or principles, and, at a more basic level, the concepts that are deemed relevant in determining the context within which a decision should be made.

Fairness should be supported by the learning process by allowing the acquired knowledge to identify possible gaps, which might lead to biased inferences, so that the learning process can be further guided to fill these gaps and resolve the biases, by seeking to identify diverse data points from which to learn, and ones that would get learning outside any filter-bubbles. Relatedly, transparency should be supported by the learning process by ensuring that learned knowledge is represented in a form and structure that is compatible with human cognition.

Argumentation can identify gaps in knowledge, and sources of potential biases, by acknowledging that individual data points can form very specific and strong arguments that defeat the general arguments based on highly-predictive features, by having arguments dispute other arguments that rely on socially or ethically inappropriate features, and by supporting dilemmas in case the evidence for and against a certain conclusion might not be fully statistically supported. In all cases, the arguments in favor and against a certain inference can be made explicit to users, so that they can deliberate, for example, on the merits of high-accuracy coming through some rules, vs. the dangers of introducing biases.

A last, by major, overarching challenge for the process of knowledge acquisition is its meaningful integration with the process of reasoning. Learned knowledge does not exist in a

vacuum, and it cannot be decoupled from how it will be reasoned with. Rather, during the learning process one has to reason with learned knowledge, so that its effects can be taken into account for the learning of further knowledge (Michael, 2014, 2016). This challenge is aligned with the challenge of learning structured and hierarchical knowledge, and the incremental nature of learning this knowledge. Once the bottom layers of knowledge are learned, they need to be used to draw intermediate inferences, so that the top layers of the knowledge can be learned to map those drawn intermediate inferences to higher inferences.

Not all layers of knowledge need to be represented as connections between identifiable concepts. At the lowest levels of learned knowledge, where inputs come in the form of unstructured (subsymbolic) data, neural architectures can play a meaningful role. As one moves from mapping those low-level inputs into identifiable concepts, one can then employ a representation that is based on symbols, enhancing the neural architecture with symbolic or cognitive layers of knowledge on top (Artur S. d'Avila Garcez, 2014; Tsamoura et al., 2021). Argumentation can take on the role of the language in which these cognitive layers of knowledge can be represented, allowing the necessary flexibility in mapping neural inputs to higher order concepts.

The developmental nature of learning, important in the context of building HCAI systems, has been studied in works on never-ending learning (Mitchell et al., 2015), curriculum learning (Bengio et al., 2009) and continual learning (De Lange et al., 2022), among others. Such works attempt to address the challenge that most current ML approaches face due to their batch-mode learning. If new data becomes available, previously trained knowledge is lost and the training process needs to start from scratch again. This process seems inefficient and improvable, in particular when we consider how humans learn over time. Mitchell et al. (2015) illustrate their suggested never-ending learning paradigm with the case of the Never-Ending Language Learner (NELL). NELL has continuously learned from the Web to read, and invents new relational predicates that extend the ontology to infer new beliefs. Bengio et al. (2009) take a different approach, what they call curriculum learning, but yet, similarly their motivation is inspired by human learning. They suggest to formalize training strategies, which define training orders, to reach faster training in the online setting and guide the training toward better regions in the parameter space to improve the overall quality of learning for deep deterministic and stochastic neural networks. Continual learning is yet another concept, where (De Lange et al., 2022) suggest to focus on artificial neural networks that can gradually extend knowledge without catastrophic forgetting.

Adopting argumentation as the target language of learning fits well with such attempts to develop continual learning processes (e.g., Michael, 2016). First, the take of argumentation on not producing definite conclusions in all cases is an explicit acknowledgment that any learned knowledge is never

complete, and that learning is a never-ending process. When new data arrives, this can lead to new arguments, which can be seamlessly integrated into existing knowledge learned from previously available data. If the new data statistically support arguments in conflict with those previously learned, the semantics of argumentation handles the conflict by producing dilemmas, without leading to the catastrophic forgetting of previously learned knowledge. In addition, these dilemmas can naturally direct, through a form of self-driven curriculum learning, the learning process to seek additional data to resolve those dilemmas.

4.3. Internal architecture

The previously described challenges of how knowledge is organized to facilitate context-sensitive inferences and at the same time is naturally acquired such that knowledge adapts across domains and time, raises the question of how this is achieved, or wired, into the human mind.

For the classification of human experience and information processing mechanisms, Newell (1990) established the four bands of cognition, consisting of the biological band, the cognitive band, the rational band and the social band. These are characterized by the timescales of twelve different orders of magnitude. As an example, the time span of processes in the cognitive band can occur in 100 ms, whereas the time span of processes within the rational band ranges from minutes to hours. Newell was probably right when stating that any theory which only covers one aspect of human behavior “flirts with trouble from the start” (Newell, 1990), and therefore he suggested the development of architectures of cognition as formal structures in which different cognitive processes can be simulated and interact as modules.

At a general level, such **Cognitive Architectures** need to provide (i) a specification of the structure of the brain, (ii) the function of the mind and (iii) how the structure explains the function (Anderson, 2007). They are required to unify different information processing structures within one system that simulates the processes organized as modular entities and that are coordinated within one environment thus simulating human cognition and eventually predict human behavior. Over the decades, many cognitive architectures such as ACT-R (Anderson, 2007) or SOAR (Laird, 2012) have been proposed, which have had a significant contribution on providing formal methodologies and have been applied to various levels of cognition by including both symbolic and subsymbolic components. Laird et al. (2017) suggest a baseline model, the ‘standard model of the mind’ (or ‘common model of cognition’), in order to ‘facilitate shared cumulative progress’ and align theories on the architectural level.

However, even after 50 years, Newell’s criticism that the scientific community does not “seem in the experimental

literature to put the results of all the experiments together” (Newell, 1973) still seems to hold. Interestingly, this missing convergence toward unified theories of cognition persists across and within the *bands of cognition* (Newell, 1990). Bridging the gap between Newell’s bands of cognition still exists as a problem and the main challenge remains. How do we organize the internal processes of a system at different levels such that they can operate internally linking perception and high-level cognition, by facilitating their meaningful integration with other systems and the external human participating environment? This is a question not only on how theories are embedded across levels, but also on which ones are adequate theories at the individual levels, and, in particular, on how organizational models are generated from theories across task domains.

The intention of HCAI to take the human perspective into account from the beginning of the system’s development, in order to support and enhance the human’s way of working, requires that its systems are judged not in terms of their optimization according to current AI performance criteria, but rather in terms of a holistic evaluation in comparison with the human mind and behavior. Laird, Lebiere and Rosenboom (Laird et al., 2017) emphasize that for human-like minds, the overall focus needs to be on ‘the bounded rationality hypothesized to be central to human cognition (Simon, 1957; Anderson, 1990)’. Accordingly, as we have stated several times in this paper, HCAI systems need to provide solutions that are not necessarily optimal in the strict rational sense but cognitively plausible across different levels. One way to address the above requirements is to build HCAI systems that have an internal representation of the current state of the human mind (Theory of Mind). This representation reflects the human’s awareness of their environment from which plausible behavior in the given context can be ascertained. The system can consider the human perspective and generate their plausible decisions, if it has the ability to simulate the human’s mind functions and their interaction with the simulated environment. Yet, the main challenge remains: How to organize the internal processes of a system at different levels such that they can operate internally in a coherent way and facilitate their meaningful integration with other systems and the external human participating environment. What is an adequate internal representation, and at which levels does the system need to be implemented? How are these levels organized internally?

Can Cognitive Argumentation help to address these challenges? Cognitive Argumentation has its foundations in Computational Argumentation and thus, at some level, its process of building arguments and the dialectic process of reasoning can be described and understood symbolically. Yet, the actual processes of building, choosing, and deciding which arguments are plausible or winners can be heavily guided by biases or heuristics which stem from lower level, e.g., statistical, components. These components might account for lower levels

of cognition such as situation awareness or associative memory. Their connection with higher-level processes, such as the relative strength relation between arguments, can thus provide a vehicle of integration between internal system processes (e.g., Dietz, 2022). Cognitive Argumentation might therefore be considered as a good candidate for the internal integration, within appropriate cognitive architectures, of the processes at different cognitive levels of HCAI systems.

4.4. Social integration

Argumentation in practice is often a social activity, carried out through a dialogue or debate among (groups of) different individuals. Similar to a multi-agent system, where independent entities are understood as agents (passive, active, or cognitive), in an argumentation environment agents can be (groups of) individuals holding to or against a certain position. Multi-agent systems in their traditional sense have been used to study the dynamics of complex systems (e.g., economic systems) and the influence of different interactive behaviors among agents. Usually, the optimal outcome is computed with respect to a rational agent's behavior, i.e., an agent who selects an action that is expected to maximize its performance measure. In the case of Human-Centric AI systems, operating in such an optimality-seeking mode is not realistic. Yet, the different systems or agents need to operate within the same environment, either in a cooperative or competitive mode, as the case may be. The important challenge for this joint and social operation is sustainability, in the sense that individual systems can continue to provide their separate services while the ecosystem in which they belong continues to support their individual roles.

How can the logical foundation of argumentation facilitate achieving this goal of social sustainability? Argumentation can be understood as a multi-agent system where each agent (or group of agents) is a representative for supporting a certain position. The overall system might contain various (groups of) agents holding to different, possibly conflicting, positions. As in multi-agent systems, such an argumentation environment can have a notion of cooperation and competition. Cooperation can be understood as agents holding to the same position, where their joint goal is to defend their position or to convince others about their position. Competition is the case where agents have opposing positions and try to defeat the other's arguments, while defending their own arguments. Interaction among these (groups of) individual systems occurs through the arguments that defend their own positions or defeat the positions of others. This then can reflect the overall system's dynamics, which might either converge toward one position or stabilize to various (strong) positions that conflict with each other.

Another view on argumentation as a multi-agent system, following the work of Mercier and Sperber (2009), is to cast one agent as a communicator and other agents as the audience.

The exchange of information happens dynamically through the persuasiveness of the communicator and the *epistemic vigilance* of the audience. In some sense this is the original context of the study of argumentation going back all the way to Aristotle who stages the process of dialectic argumentation between a Questionnaire and an Answerer. The motivation is to understand how to regulate the process of communication, e.g., exposing unreasonable positions and harmful rhetoric. In today's explosion of media and *social networks* this is particularly important in helping to enhance the quality of dialogue and interaction on these platforms (Heras et al., 2013; Gurevych et al., 2017). Recently, the center of Argument Technology (<https://arg-tech.org/>) has released a video exposing the dangers of harmful rhetoric, arguing that argumentation technology can help address this problem, e.g., with systems that support "reason checking" of the premises and validity of a position promoted on the media and social cyberspace.

In all cases, the approach needs to be strongly guided by cognitive heuristics (e.g., 'bias by authority', or heuristics concerned with the ethical aspects). The overall major challenge then remains the same. How can HCAI systems be socially integrated within an application environment for dialogue and debates? How can argumentation and the argumentative structure of knowledge facilitate such an integration?

4.5. Ethical compliance

The ethical requirement of HCAI systems is of paramount and unique importance. Its importance is reflected by the unprecedented interest and proactive actions that organizations and governments are taking in order to safeguard against possible unethical effects that AI can have on people's lives.¹⁰

One such EU initiative is the publication of "Ethics Guidelines for Trustworthy AI"¹¹, prepared by a "High-Level Expert Group on AI," suggesting that AI systems should conform to seven different requirements in order to be ethical and trustworthy (see also Floridi, 2014; Russell, 2019). At the systemic operational level, one of these requirements is that of the "Transparency: Including traceability, explainability and communication" of the system. This requirement alludes to the importance of AI systems being able to enter into a dialogue and a debate with human users or other such systems, and for this to be meaningful the system should be able to explain and account for its decisions and position. This will ensure some level of ethical behavior as through these processes of dialogue, dispute,

¹⁰ The EU is continuously releasing documents of guidelines and regulatory or legal frameworks on AI Ethics, e.g., <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence>.

¹¹ https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guideline_s.pdf

and debate we will be able to identify ethical weaknesses and take action to remedy or mitigate the problem. The challenge then for any logical foundation of AI is to facilitate these processes and allow in a modular and natural way the adaptation of the systems with the results of the debate, either at the level of its knowledge, or at the level of its internal operation.

Transparency and other such requirements provide an operational approach to the problem. They do not touch, though, on the underlying foundational difficulty of what is good ethical behavior and how we can endow AI systems with it. The inherent difficulty in achieving the, otherwise simply stated, challenge of “AI systems that adhere to human moral values” lies in the fact that even if we are clear about the moral values by which we generally want to regulate our systems, in many circumstances we might have different moral values that are in conflict with each other.

The problem is not new. It is as old as Philosophy, where it was recognized that within ethical reasoning we can often have **moral dilemmas** of being unable to decide clearly what is the correct ethical decision or action to take. Socrates from the very early days of Philosophy raises this concern of morally difficult and unclear decisions depending on the particular context at hand, and Aristotle aims to give prescriptions for ethical reasoning in his Practical Syllogisms. Recently, in the context of AI, the Moral Machine project (Bonnefon et al., 2016) draws from the *miners dilemma* in Philosophy, in an attempt to gather data on the moral values of people and the relative importance they place on them, albeit within a very specific “AI context” that is directly relevant to the increasing prevalence of autonomous cars.¹² The project confirms that decisions in ethical reasoning are not always clear and that they can vary between different people.

From this theoretical point of view it appears that the essential difficulty in this challenge for ethical decisions is that of capturing the context-sensitive nature of the reasoning involved. This is, therefore, the same problem described in Sections 4.1 and 4.2, where we have considered the nature of reasoning and learning in Human-Centric AI systems.

The flexibility of the Logic of Argumentation is well suited for the ethical guidelines, which although strong, they cannot be absolute, as situations can arise with genuine moral dilemmas (Verheij, 2016).¹³

In general, as we consider the challenge of how to develop the ethical quality within our AI systems, it would be useful to be able to judge the current degree of achieving this, i.e., what we could call the current level of **ethicacy** of a system.¹⁴

¹² <https://www.moralmachine.net/>

¹³ Also consider https://www.ai.rug.nl/~verheij/publications/oratie/oratie_Bart_Verheij.pdf, <https://www.argnet.org/ethics-of-arg>.

¹⁴ Ethicacy: the efficacy in achieving ethical behavior; a measure of the ability to operate ethically to a satisfactory or expected degree.

The form that this ethicacy measure would have depends on the logical perspective that we adopt about the ethical requirements, e.g., whether these are normative directives or guidelines to follow based on some descriptive principles. The normative view would point toward “ethics by design,” whereas the descriptive view would point toward an “evolutionary process.” Adopting the more flexible descriptive perspective, as argumentation would allow — instead of appealing to either ethics experts to prescribe, or supervised learning techniques to induce, the ethical principles — can support also a process of gradual acquisition of these principles. This process would resemble how young children learn from their parents and social surroundings: by being coached in an online and developmental manner as a reaction to their ethical transgressions (Michael, 2019, 2020).

Such a process of “ethics coaching,” be it by the user being assisted by the system, or by ethics experts acting on behalf of some community, or indeed special Ethics Coaching AI systems, can react to contest the decision of the system and possibly help to resolve the dilemma under some specified conditions. Critical in this interaction is that it is the justifications being evaluated, and not only the inferred conclusion, and that the reaction comes in the form of ethical counterarguments that do not completely nullify the system’s current ethical principles, but complement them in an elaboration tolerant manner. Hence the ethical dimension of a system can start with some, pre-populated by design (by ethics experts) broad generally-accepted, ethical principles to guarantee some minimally-viable version of the system. Then, every time the system is faced with an ethically-driven dilemma on its material choices, the ethics coaching process will help the system, through a coaching dialogue on the justification of the alternatives, develop higher levels of ethicacy.

Argumentation, as a logical foundation supporting an ethical behavior, would allow machines to make transparent the reasons in favor and against the options available, and make transparent the ways in which these reasons are further developed and refined over time. Exposing the reasoning in one’s decisions would seem to be the primary desideratum for an ethical system, over and above what the actual decision might end up being. At the end of the day, different people (or a system and a user) might disagree on their ethical principles. At the very least, argumentation can help expose the fundamental premises on which interlocutors disagree, even if it cannot help them reconcile their divergent views.

In his inaugural lecture,¹⁵ Verheij proposes not to regulate AI by enforcing human control or by the prohibition of ‘killer robots,’ but through the use of argumentation systems

¹⁵ https://www.ai.rug.nl/~verheij/publications/oratie/oratie_Bart_Verheij.pdf

which provide us with good arguments. *You cannot force good ethical behavior, you can only hope that you can form such behavior through exposure to the arguments for the alternatives.*

4.6. Summary of HCAI challenges

We can summarize these challenges by regrouping them into three main groups of different type of “technical requirements” expected from the logical foundations of HCAI systems and connecting to each one of these the main feature of argumentation that is appropriate to meet these requirements. Table 2 shows these three groups of requirements: **Openness** to capture the open nature of operation and development of the systems, **Humanly** to give the systems a human-like compatible behavior, and **Ethicacy** to capture the need for these systems to be effectively regulated by human moral values. In the second column of this table, we have the main corresponding features of argumentation that can help in addressing these requirements: The **flexible and non-strict** nature of argumentative logical inference together with the **online** process of argumentation are directly relevant in addressing the needs of the first group. For the second group of the requirements we note that the inference of argumentation is naturally **human-like**: human cognition and reasoning is naturally carried out through argumentation. The **dialectic** process of argumentation occurs in a framework of inner contemplation or debate between alternative points of view. This together with the natural link of arguments as justifications or explanations for supporting a view against

others, can form the basis on which to build the required processes in the third group of the ethicacy requirements.

5. Conclusions

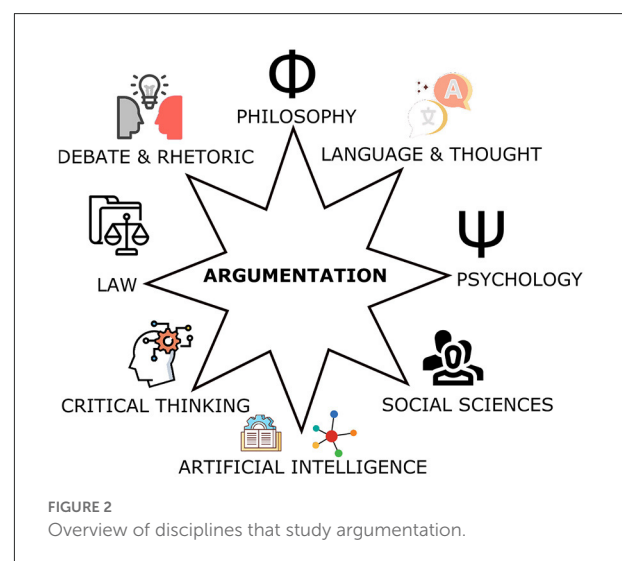
We have proposed Argumentation as a candidate for the logical foundations of Human-Centric AI. This position is based on the close and natural link of argumentation with human cognition. Argumentation as a formal system of reasoning could provide the underlying framework for computational models of human-like intelligent faculties for AI systems. The overall idea is that by allowing machines to argue, and by bringing their form of argumentation close to human argumentation, we can facilitate a smooth machine-human interaction that offers an enhancement of people’s general intelligent capabilities in a natural way that is ethical and humane.

Whatever logic we choose, and no matter how appropriate we judge it to be, as a logical foundation for HCAI, this can only be the first step toward developing HCAI systems. Intelligence, whether human or artificial, is not a matter of pure logic as we are reminded by Kant and McDermott in their works “Critique of Pure Reason” (McDermott, 1990; Kant, 1998). A logical foundation needs to enable and facilitate the use of extra-logical cognitive information (or cognitive principles), in order to turn the underlying reasoning and learning that are supported by the logic into cognitive processes. Logic is not applied in isolation, but needs to be “aware” of a cognitive operational framework that affects and regulates its application. This cognitive embodiment would require the synthesis of knowledge from a wide range of disciplines that study the different aspects of human thought in its full generality.

We are thus presented with an additional epistemological challenge, on top of the other technical challenges, of addressing the need for an interdisciplinary synthesis of the various studies

TABLE 2 Summary of technical challenges of HCAI, expected to be supported by its logical foundations and appropriate general properties of argumentation.

HCAI Technical Challenges		Argumentation Properties
Openness	Context-sensitive inference	Flexibility of argumentation logic
	Online and adaptive inference	
	Continuous and adaptive learning	
	Tolerance of inference to incompleteness and conflicting information	
Humanly	Cognitively compatible system-human interaction	Argumentation-based human cognition
	Personalization of inference	
	Responsiveness to users feedback	
	Socially-driven inference	
Ethicacy	Cognitive explainability and transparency	Dialectic nature of argumentation
	Contestable dialogues and debates	
	Corrective moral/ethical coaching	
	Osmotic learning of ethical behavior	



of human argumentation under the perspective of Human-Centric AI. How can we draw from these different fields to form a foundation where machine argumentation is brought cognitively close to human argumentation? What empirical studies of human intelligence in these fields will help us understand its link with machine intelligence and particularly with computational argumentation, in a way useful for building HCAI systems? What elements of these fields are needed to allow the development of Human-Centric AI as a truly interdisciplinary field? For the case of argumentation, we are fortunate to have a wide ranging study of argumentation within several disciplines, such as Cognitive Psychology, Critical Thinking, Debate and Rhetoric, Argumentative Discourse in Natural Language, and studies of Practical Argumentation in different human contexts (see Figure 2). We can then draw from these studies to help us in addressing the interdisciplinary nature of HCAI.

Ideally, we would want this interdisciplinary synthesis to be so strong that Human-Centric AI would generate feedback into these other disciplines and become itself part of the general effort to understand human thought and intelligence. Can Human-Centric AI give a focus for pulling together the different efforts to comprehend human intelligence, and function as a new “laboratory space” for evaluating and further developing our understanding of the many different facets of human thought?

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

References

- Anderson, J. R. (1990). *The Adaptive Character of Thought*. New York, NY: Psychology Press.
- Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford: Oxford University Press.
- Artur, S., d'Avila, G., Dov, M., and Gabbay, L. C. L. (2014). A neural cognitive model of argumentation with application to legal inference and decision making. *J. Appl. Logic* 12, 109–127. doi: 10.1016/j.jal.2013.08.004
- Atkinson, K., Baroni, P., Giacomini, M., Hunter, A., Prakken, H., Reed, C., et al. (2017). Towards artificial argumentation. *AI Mag.* 38, 25–36. doi: 10.1609/aimag.v38i3.2704
- Bench-Capon, T. J. M., and Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artif. Intell.* 171, 619–641. doi: 10.1016/j.artint.2007.05.001
- Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). “Curriculum learning,” in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09* (New York, NY: Association for Computing Machinery), 41–48.
- Bondarenko, A., Dung, P. M., Kowalski, R. A., and Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. *Artif. Intell.* 93:63–101. doi: 10.1016/S0004-3702(97)00015-5
- Bonet, B., and Geffner, H. (2001). Planning as heuristic search. *Artif. Intell.* 129, 5–33. doi: 10.1016/S0004-3702(01)00108-4

Funding

This work was supported by funding from the EU's Horizon 2020 Research and Innovation Programme under grant agreements (nos. 739578 and 823783) and from the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation, and Digital Policy.

Acknowledgments

The authors gratefully acknowledge the valuable feedback by the reviewers. AK wishes to acknowledge the collaboration with Doctors Vasilis and Panayiotis Tanos on the development of a real-life system, called GAID: Gynecology AI Diagnostics, helping him to understand the challenges of AI in practice.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science* 352, 1573–1576. doi: 10.1126/science.aaf2654
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press.
- Buchanan, B. G., and Shortliffe, E. H. (Eds.). (1984). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley.
- Cocarascu, O., and Toni, F. (2016). “Detecting deceptive reviews using argumentation,” in *Proceedings of the 1st International Workshop on AI for Privacy and Security, PrAISe@ECAI 2016, The Hague, Netherlands, August 29–30, 2016* (Hague: ACM), 9:1–9:8.
- Cyras, K., Rago, A., Albin, E., Baroni, P., and Toni, F. (2021). “Argumentative xai: a survey,” in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, ed Z.-H. Zhou (Montreal: International Joint Conferences on Artificial Intelligence Organization), 4392–4399.
- Dagan, I., Dolan, B., Magnini, B., and Roth, D. (2009). Recognizing textual entailment: rational, evaluation and approaches. *Natural Lang. Eng.* 15, I-Xvii. doi: 10.1017/S1351324909990209
- De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., et al. (2022). A continual learning survey: defying forgetting in classification tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 3366–3385. doi: 10.1109/TPAMI.2021.3057446
- Dietz, E. (2022). “Argumentation-based reasoning guided by chunk activation in ACT-R,” in *Proceedings of the 20th International Conference on Cognitive Modelling*.
- Dietz, E., and Kakas, A. C. (2021). “Cognitive argumentation and the selection task,” in *Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 43* (Cognitive Science Society), 1588–1594.
- Evans, J. S. B. T. (2010). *Thinking Twice: Two Minds in One Brain*. Oxford: Oxford University Press.
- Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford: OUP Oxford.
- Gracia, A. J., and Simari, G. R. (2004). Defeasible logic programming: an argumentative approach. *Theory Pract. Logic Program.* 4, 95–138. doi: 10.1017/S1471068403001674
- Gurevych, I., Lippi, M., and Torroni, P. (2017). Argumentation in social media. *ACM Trans. Internet Techn.* 17, 1–2. doi: 10.1145/3056539
- Halpern, J. Y., Harper, R., Vardi, M. Y., and Immerman, N. (2001). On the unusual effectiveness of logic in computer science. *Bull. Symbolic Logic* 7, 1–19. doi: 10.2307/2687775
- Heras, S., Atkinson, K., Botti, V. J., Grasso, F., Julián, V., and McBurney, P. (2013). Research opportunities for argumentation in social networks. *Artif. Intell. Rev.*, 39, 39–62. doi: 10.1007/s10462-012-9389-0
- Hinton, M. (2019). Language and argument: a review of the field. *Res. Lang. Łódź* 17, 93–103. doi: 10.2478/rela-2019-0007
- Hinton, M. (2021). *Evaluating the Language of Argument*. Łódź: Springer Nature Switzerland AG.
- Horty, J. F., Thomason, R. H., and Touretzky, D. S. (1990). A skeptical theory of inheritance in nonmonotonic semantic networks. *Artif. Intell.* 42, 311–348. doi: 10.1016/0004-3702(90)90057-7
- Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.
- Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291. doi: 10.2307/1914185
- Kakas, A., and Michael, L. (2020). Abduction and argumentation for explainable machine learning: A position survey. *arXiv [Preprint]*. arXiv: 2010.12896. doi: 10.48550/ARXIV.2010.12896
- Kakas, A. C. (2019). Informalizing formal logic. *Informal Logic* 39, 169–204. doi: 10.22329/il.v39i2.5169
- Kakas, A. C., Mancarella, P., and Toni, F. (2018). On argumentation logic and propositional logic. *Studia Logica* 106, 237–279. doi: 10.1007/s11225-017-9736-x
- Kakas, A. C., and Moraitis, P. (2003). “Argumentation based decision making for autonomous agents,” in *Proceedings of 2nd International Joint Conference on Autonomous on Autonomous Agents and Multiagent Systems, AAMAS* (Melbourne: ACM), 883–890.
- Kant, I. (1998). *Critique of Pure Reason. The Cambridge Edition of the Works of Immanuel Kant*. New York, NY: Cambridge University Press. Translated by Paul Guyer and Allen W. Wood.
- Laird, J. E. (2012). *The Soar Cognitive Architecture*. Cambridge, MA; London: The MIT Press.
- Laird, J. E., Lebiere, C., and Rosenbloom, P. S. (2017). A standard model of the mind: toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Mag.* 38, 13. doi: 10.1609/aimag.v38i4.2744
- Lawrence, J., and Reed, C. (2019). Argument mining: a survey. *Comput. Linguist.* 45, 765–818. doi: 10.1162/coli_a_00364
- Lieto, A. (2021). *Cognitive Design for Artificial Minds*. Abingdon; New York, NY: Routledge.
- Lippi, M., and Torroni, P. (2016). Argumentation mining: state of the art and emerging trends. *ACM Trans. Internet Techn.* 16, 1–10. doi: 10.1145/2850417
- Marek, W., and Truszczyński, M. (1991). Autoepistemic logic. *J. ACM* 38, 587–618. doi: 10.1145/116825.116836
- McCarthy, J. (1968). “Programs with common sense,” in *Semantic Information Processing* (MIT Press), 403–418.
- McDermott, D. (1990). A critique of pure reason 1. *Comput. Intell.* 3, 151–160. doi: 10.1111/j.1467-8640.1987.tb00183.x
- Mercier, H., and Sperber, D. (2009). “Intuitive and reflective inferences,” in *Two Minds: Dual Processes and Beyond* (Oxford), 149–170.
- Mercier, H., and Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behav. Brain Sci.* 34, 57–74. doi: 10.1017/S0140525X10000968
- Michael, L. (2014). “Simultaneous learning and prediction,” in *Proceedings of the 14th International Conference on the Principles of Knowledge Representation and Reasoning*, Vienna.
- Michael, L. (2016). “Cognitive reasoning and learning mechanisms,” in *Proceedings of the 4th International Workshop on Artificial Intelligence and Cognition* (New York, NY), 2–23.
- Michael, L. (2019). “Machine coaching,” in *Proceedings of IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)* (Macao), 80–86.
- Michael, L. (2020). “Machine ethics through machine coaching,” in *Proceedings of 2nd Workshop on Implementing Machine Ethics @ UCD*, (Dublin).
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267:1–38. doi: 10.1016/j.artint.2018.07.007
- Minsky, M. (1981). “A framework for representing knowledge,” in *Mind Design: Philosophy, Psychology, Artificial Intelligence*, ed J. Haugeland (Cambridge, MA: MIT Press), 95–128.
- Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., et al. (2015). “Never-ending learning,” in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, (Palo Alto, CA).
- Modgil, S., and Prakken, H. (2013). A general account of argumentation with preferences. *Artif. Intell.* 195, 361–397. doi: 10.1016/j.artint.2012.10.008
- Newell, A. (1973). “You can’t play 20 questions with nature and win: Projective comments on the papers of this symposium,” in *Visual information* (New York, NY: Academic Press).
- Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press, USA.
- Nisbet, R., Miner, G., and Yale, K. (Eds.). (2018). *Handbook of Statistical Analysis and Data Mining Applications, 2nd Edn*. Boston, MA: Academic Press.
- Paglieri, F., and Castelfranchi, C. (2010). Why argue? towards a cost-benefit analysis of argumentation. *Argument Comput.* 1, 71–91. doi: 10.1080/19462160903494584
- Perelman, C., and Olbrechts-Tyteca, L. (1969). *The New Rhetoric. A Treatise on Argumentation*. Notre Dame; London: University of Notre Dame Press.
- Pollock, J. L. (1987). Defeasible reasoning. *Cogn. Sci.* 11, 481–518. doi: 10.1207/s15516709cog1104_4
- Prakken, H. (2010). An abstract framework for argumentation with structured arguments. *Argument Comput.* 1, 93–124. doi: 10.1080/19462160903564592
- Reiter, R. (1980). A logic for default reasoning. *Artif. Intell.* 13, 81–132. doi: 10.1016/0004-3702(80)90014-4
- Russell, S. J. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, NY: Penguin Publishing Group.
- Saldanha, E. D., and Kakas, A. C. (2019). Cognitive argumentation for human syllogistic reasoning. *Künstliche Intell.* 33, 229–242. doi: 10.1007/s13218-019-00608-y

- Schank, R. C., and Abelson, R. P. (1975). "Scripts, plans and knowledge," in *Thinking: Readings in Cognitive Science, Proceedings of the Fourth International Joint Conference on Artificial Intelligence*, eds P. Johnson-Laird and P. Wason (Tbilisi, GA: USSR), 151–157.
- Shoham, Y. (1987). "Nonmonotonic logics: meaning and utility," in *Proceedings of 10th International Joint Conferences Artificial Intelligence (IJCAI-87)* (Milan), 388–393.
- Simari, G. R., and Rahwan, I. (Eds.). (2009). *Argumentation in Artificial Intelligence*. New York, NY: Springer.
- Simon, H. A. (1957). "A behavioral model of rational choice," in *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting* (New York, NY: John Wiley and Sons).
- Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.
- Tsamoura, E., Hospedales, T., and Michael, L. (2021). Neural-symbolic integration: a compositional perspective. *Proc. AAAI Conf. Artif. Intell.* 35, 5051–5060. doi: 10.1609/aaai.v35i6.16639
- Turing, A. M. (1947). *Lecture to London Mathematical Society, February 20, 1947*. Turing Digital Archive.
- van Eemeren, F. H., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., Verheij, B., and Wagemans, J. H. M. (2014). *Handbook of Argumentation Theory, 1st Edn.* Dordrecht: Springer Netherlands; Imprint; Springer.
- van Eemeren, F. H., and Grootendorst, R. (2004). *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*. Cambridge: Cambridge University Press.
- Vassiliades, A., Bassiliades, N., and Patkos, T. (2021). Argumentation and explainable artificial intelligence: a survey. *Knowl. Eng. Rev.* 36, 11. doi: 10.1017/S0269888921000011
- Verheij, B. (2016). Formalizing value-guided argumentation for ethical systems design. *Artif. Intell. Law* 24, 387–407. doi: 10.1007/s10506-016-9189-y
- Wagemans, J. H. M. (2018). Analogy, similarity, and the periodic table of arguments. *Stud. Logic Grammar Rhetoric* 55, 63–75. doi: 10.2478/slgr-2018-0028
- Walton, D. (1996). *Argumentation Schemes for Presumptive Reasoning*. Mahwah, NY: Psychology Press.
- Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge: Cambridge University Press.



OPEN ACCESS

EDITED BY
Antonios Kakas,
University of Cyprus, Cyprus

REVIEWED BY
Marcos Cramer,
Technical University Dresden, Germany
Yiannis Kiourekis,
University of Thessaly, Greece

*CORRESPONDENCE
Kaan Kilic
✉ kaank@cs.umu.se

SPECIALTY SECTION
This article was submitted to
Machine Learning and Artificial Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

RECEIVED 13 October 2022
ACCEPTED 20 January 2023
PUBLISHED 16 February 2023

CITATION
Kilic K, Weck S, Kampik T and Lindgren H (2023)
Argument-based human–AI collaboration for
supporting behavior change to improve health.
Front. Artif. Intell. 6:1069455.
doi: 10.3389/frai.2023.1069455

COPYRIGHT
© 2023 Kilic, Weck, Kampik and Lindgren. This
is an open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Argument-based human–AI collaboration for supporting behavior change to improve health

Kaan Kilic*, Saskia Weck, Timotheus Kampik and Helena Lindgren

Department of Computing Science, Umeå University, Umeå, Sweden

This article presents an empirical requirement elicitation study for an argumentation-based digital companion for supporting behavior change, whose ultimate goal is the promotion and facilitation of healthy behavior. The study was conducted with non-expert users as well as with health experts and was in part supported by the development of prototypes. It focuses on human-centric aspects, in particular user motivations, as well as on expectations and perceptions regarding the role and interaction behavior of a digital companion. Based on the results of the study, a framework for person tailoring the agent's roles and behaviors, and argumentation schemes are proposed. The results indicate that the extent to which a digital companion argumentatively challenges or supports a user's attitudes and chosen behavior and how assertive and provocative the companion is may have a substantial and individualized effect on user acceptance, as well as on the effects of interacting with the digital companion. More broadly, the results shed some initial light on the perception of users and domain experts of "soft," meta-level aspects of argumentative dialogue, indicating potential for future research.

KEYWORDS

formal argumentation dialogues, behavior change, digital companion, value-based argumentation, argumentation schemes, user-modeling, Human-Centered Artificial Intelligence, health promotion

1. Introduction

Artificially intelligent agents in the form of digital assistants, or companions (Torous et al., 2018), are to an increasing extent being developed for supporting individuals with improving health by changing unhealthy behavior. However, each individual has different motives for attempting a change of behavior and different reasons for why they do not achieve the desired behavior. These motives and reasons can be formulated as arguments, which can potentially be used as the basis for argument-based dialogues between an individual and a digital companion. Moreover, users may have different perceptions of how an agent could collaborate and provide support in the process, which may affect how argument-based dialogues with a digital companion can unfold.

Although there are plenty of examples of behavior change support applications, few apply computational argumentation frameworks as the foundation for organizing motives in favor and against what to do to promote health and in the reasoning in deliberative dialogues between the human and a digital agent.

The purpose of the research presented in this article is to explore from a user's perspective and from the perspectives of experts on behavior change a digital companion with which the user can have argument-based dialogues in the process of behavior change, which the user can tailor to adhere to their expectations regarding roles and types of support in the dialogues. The aim is to provide the user with means to collaborate with the digital agent to ultimately become empowered and supported in their pursue of their goals to improve their health. Research presented, in this article, is consequently an example of Human-Centered Artificial Intelligence (HCAI), which is defined by Nowak et al. (2018) as AI that collaborates with a human, "enhancing their capabilities, and empowering them to better achieve their goals."

Our study explores the following research questions:

- What are people's expectations of a digital coach or companion in terms of roles and behaviors, and argument-based support?
- How can the agent's roles and behavior, and the argument-based dialogue promoting health be tailored to individuals' expectations and level of readiness for the change?

The main contributions to the field of HCAI are (i) increased knowledge about how people view argument-based support through digital companions for promoting healthy lifestyles, (ii) an argumentation-based framework for tailoring a digital agent's roles and behaviors, and (iii) a novel application of argumentation schemes for tailoring a digital companion's role and behavior and for switching between or merging roles. The article exemplifies how computational argumentation provides the foundation for HCAI for supporting behavior change to improve health.

The remainder of the article is organized as follows. First, the conducted research is contextualized and an overview of related work in computational argumentation and human-computer interaction is provided in section 2. Next, the methodology applied in the studies conducted is presented in section 3. The results are provided in section 4 and are discussed in section 5. Conclusion is provided in section 6.

2. Background and related work

The research presented in this article is conducted as a part of a research project exploring digital companions as social actors related to managing stress, and the research program STAR-C, which aims to develop a digital coach for promoting healthy lifestyle habits targeting physical activity, nutrition, alcohol consumption, tobacco use, and stress (Lindgren et al., 2020; Ng et al., 2021). The STAR-C program builds on and extends the Västerbotten Health Intervention Program (VIP) in which the population in the healthcare regions are invited to a health checkup that includes motivational interviewing with a specially trained nurses when turning 40, 50, and 60 years old (Hörnsten et al., 2014). The VIP is successful in reducing premature cardiovascular disease mortality and extending a healthy life in a cost-effective manner and has become a health promotion model also for other regions (Blomstedt et al., 2015; Lindholm et al., 2018).

The concept of digital companions for maintaining a healthy lifestyle and goal achievement is increasingly gaining attention; it is, for example, applied and studied in the context of professional work support, education, stress management, healthcare, and behavior

change (Jang and Kim, 2020; Braun et al., 2021; Spirig et al., 2021; Weber et al., 2021; Manning et al., 2022). All the facets, however, converge on similar topics, such as assessing the user's context or learning more about the user's habits in the interest of providing personalized support to address a specific problem. Such knowledge about the user is then embedded in a user model, which guides the system in tailoring its behavior to an individual's needs and preferences (Kobsa, 1990). Increasingly, the importance of building an artificial theory of mind (ToM) in digital and robotic companions similar to what humans do about others in order to understand and predict others' behaviors and intentions, has been pointed out, recently as being one of three grand challenges for human-AI interaction (Yang et al., 2018) that is instrumental to human-centered AI (Nowak et al., 2018). To achieve this, models are required that integrate different aspects such as episodic memory, empathy, hierarchical models of activity, and tasks to advance the capabilities (Steels, 2020).

The person-tailored argumentation-based decision-support system developed as a part of this research rests on complementary theoretical frameworks developed within different fields of research to encompass the human-centric approach: (i) on human activity (Kaptelinin and Nardi, 2006; Lindgren and Weck, 2022), (ii) motivation and behavior change (Ryan and Deci, 2000; Prochaska et al., 2015; Lindgren and Weck, 2021), (iii) argumentation theory (Walton and Krabbe, 1995; Bench-Capon, 2002; Walton et al., 2008), and (iv) formal argument-based dialogues for reasoning about health (Atkinson et al., 2006; Lindgren et al., 2020; Guerrero and Lindgren, 2021a,b).

Goal setting is one of the most important personalization feature for promoting behavior change (op den Akker et al., 2014). Using goal setting along with feedback for motivational effectiveness is a very simple yet potent approach to induce a sense of accomplishment and behavior change in users (Locke and Latham, 1984; Lunenburg, 2011). It also leads to a better performance in the attempts to complete the goals and gain motivation (Latham and Locke, 1991). According to Locke and Latham (1984), introducing challenging but specific and achievable goals lead to clearer expectations of what a person must do for behavior change. According to Ryan and Deci (2000), motivation is "to be moved to do something" and a need for *autonomy*, *competence* and *relatedness* are the attributes that need to be satisfied in order to bring about intrinsic motivation in a person or, possibly, cause an orientation shift in those who were initially not intrinsically motivated. Internalization and the accommodation of the three attributes of motivation are important for user acceptance, sustainable behavior change, and obtaining goal commitment, which are heavily related to contextual and informed feedback communicated to the individual (Locke and Latham, 1984; Ryan and Deci, 2000; Jang and Kim, 2020).

Activity theory guides in this study the organization of arguments based on their content, in addition to providing the framework for understanding the human in interaction with AI systems. Activity theory defines purposeful human activity as being directed by a *motive*, responding to a human's underlying *needs* (Kaptelinin and Nardi, 2006), and composed of an hierarchy of goal-directed *actions*. At the lowest level, the *operational* tasks are found, those that are internalized and conducted without cognitive effort. Large part of a human's habits are governed at this level, without consciously deliberating on why or how to do a particular task (walking, taking the elevator instead of the stairs, sitting down, taking the car to

work, etc.). In setting goals and deliberating on what to do to promote healthy habits, e.g., in motivational dialogues with a nurse or in argument-based dialogues with a digital companion, moving between the levels of the activity hierarchy is necessary to find the grounds for why doing a particular action or activity, to formulate the motivating arguments relevant and importance for the individual. The connection between needs, long-term goals, and short-term goals was explored by Lindgren and Weck (2022), and a model of activity encompassing the building blocks for arguments across the levels of activity was defined. Furthermore, to identify the factors affecting an individual's motivation to change behavior, a model of the behavior change progress was built based on the most influential theories on motivation and behavior change (Lindgren and Weck, 2021). These two models build the basis for a user model, or ToM, for the digital companion to use in dialogues with the individual in this study.

Argumentation theory and its application in machine reasoning is an established research field encompassing formal frameworks for constructing, analyzing, and evaluating arguments, typically organized in argumentative dialogues for different purposes, e.g., for generating new knowledge, deliberating on what to do, or to persuade another agent (Walton and Krabbe, 1995). A notable foundational work on computational argumentation is Dung's study on abstract argumentation, in which arguments and conflicts between them are modeled as directed graphs—so-called argumentation frameworks (Dung, 1995).

In order to embed various factors affecting natural dialogues, formal frameworks have been developed which handle values (Bench-Capon, 2002), preferences (Amgoud and Cayrol, 2013), and audiences (Bench-Capon et al., 2007). Bench-Capon (2002) introduced *value-based* argumentation frameworks by adding a set of values that can be associated with arguments. The idea in using value-based argumentation was to have attacks between arguments failing or succeeding based on the importance of certain values that are referenced by conflicting arguments. Traditionally, computational argumentation has been a primarily formal field of study, but recently, its potential for facilitating human-machine interaction has led to increasingly applied for work, notably in the context of explainable AI (Čyras et al., 2021; Vassiliades et al., 2021) and persuasive technologies (Hadoux et al., 2018; Donadello et al., 2022). Beyond that, researchers have started to ask foundational questions about the integration of formal argumentation with cognitive perspectives, e.g., to study to what extent non-experts find the behavior of different abstract argumentation semantics intuitive (Guillaume et al., 2022) and to model “extra-logical” cognitive reasoning (i.e., reasoning that may be considered irrational from a classical logic point of view) using formal means (Dietz and Kakas, 2021).

Although there are plenty of examples of behavior change support applications, few apply computational argumentation frameworks as a foundation for organizing motives in favor and against what to do to promote health, and in the reasoning in deliberative dialogues between the human and a digital agent. Among the few examples that have used argumentation frameworks for behavior change, an early example in the nutrition domain is provided by Grasso et al. (2000), who explored dialectical argumentation embedding the transtheoretical model of change (TTM) (Prochaska et al., 2015). De Boni et al. (2006) used argumentation through a therapy system in order to change behavior in exercise. Their goal

was to apply their system to a specific issue in exercise behavior and to assess the automation capabilities of this system in future studies by improving the argumentation capabilities of the system through personalizing the language used while conversing with the client. Baskar et al. (2017) explored multipurpose argument-based dialogues through a team of agents taking on different roles pursuing different goals in order to address an individual's various sometimes conflicting motives. Roles and an agent's arguments were connected to *argumentation schemes* (Walton et al., 2008), to provide weight on how reliable the argument may be based on the source of the argument.

Chalaguine et al. (2019) and Hadoux and Hunter (2019) investigated how the concerns of the users affect the strength of arguments in dialogue, similar to Baskar et al. (2017). For instance, a user who is not too interested in, say, quitting smoking might become interested if the persuader suggests improvements that quitting can bring out in other aspects of life that the user is more inclined toward, such as social relations and physical activity. Some individuals are more predisposed to act based on their values rather than persuasion through facts (Chalaguine et al., 2019). Atkinson and Wyner (2013) define values as “social interests that a person/agent wishes to promote.” Values are relatively scalable to other values and are important for digital companions in helping a user achieve their goals because values describe desirable goals people want to achieve (van der Weide, 2011). In fact, Perelman and Olbrechts-Tyteca (1969) outlined how people do not use facts but rather their opponents' values and opinions to justify their argument.

The complementary roles of a team of digital coaches to support an individual were outlined by Baskar et al. (2017) for the purpose of managing potentially conflicting motives and needs. A similar approach is presented by Kantharaju et al. (2019); the authors integrate argumentation in a virtual multi-coach platform, in which a group of multiple coaches with their own respective field of expertise and behaviors jointly try to promote healthy behavior in a user. In their study, the authors relate their work to the argumentation schemes *Argument from Expert Opinion* (Walton et al., 2008), and their method of presenting these arguments is implemented through a dialogue game building platform. Some key challenges are listed such as differences in users and how their multi-coach platform can overcome disagreements between the virtual coaches themselves. Kantharaju et al. (2019) also delve into the usage of persuasive social agents for behavior change and which action should be taken by the virtual coaches based on success or failure in abstract argumentation.

Another approach undertaken was by Nguyen and Masthoff where they directed their focus on the effectiveness of motivational interviewing (MI) as opposed to argumentation to persuade the users in their study (Nguyen and Masthoff, 2008). They found that, in some instances, MI is more persuasive than argumentation and that the difference between tailored and non-tailored persuasive dialogue systems are negligible. Miller and Rollnick (2012) described MI as “using a person's own reasons for change within an atmosphere of acceptance and compassion.” The use of MI was also studied by Hörnsten et al. (2014), where the primary healthcare nurses use MI during their health dialogues with patients in order to have a richer and empathy building communication. Hörnsten et al. (2014) conducted 10 interviews with the primary healthcare nurses in the VIP and studied their strategies in their dialogues. Several main themes arose after the interviews, such as “guiding vs. pressuring

patients,” “adjusting vs. directing the conversation with the patients” to “inspiring confidence vs. instilling fear.” It is concluded in their study that patient-centered care is preferable, and one key finding in the study is that ideal consultations between the nurse and the patient require empowering words, whereas consultations that include a non-willing patient for behavior change might necessitate pressure, demands for responsibility and challenge.

The need for both supportive and challenging arguments for increasing motivation suggests that a *bi-polar* argumentation framework is suitable to capture both the aspects of challenging the human to change behavior using arguments on the one hand, while also embedding the advantages of MI’s sense of acceptance and compassion on the other hand. A bi-polar argumentation framework embeds both arguments in favor and against, for instance, an activity to be conducted (Amgoud et al., 2008). Furthermore, embedding values representing the strength of an argument would allow for comparing arguments (Bench-Capon, 2002). While the atmosphere of acceptance and compassion may be promoted by providing supporting arguments, an emotional parameter expressed as friendliness or empathy is typically expected in inter-human dialogues and has been shown to be also expected in human–robot dialogues, e.g., by Tewari and Lindgren (2022).

To summarize, one of the challenges of this study is to acknowledge the ethical concerns related to evoking cognitive dissonance and potential fear in the individual when challenging their unhealthy choices on the one hand, and on the other hand, providing acceptance and compassion as in MI. The unavoidable human emotional component of arguments and argumentation relating to an individual’s choices affecting health is in the following addressed by eliciting the user’s preferences regarding the agent’s behavior. These preferences are treated as agreements between the user and the agent on how the user expects the agent to perform argument-based dialogues and can be considered a kind of social norm.

3. Methods

The research presented in this article applies a constructive, participatory design methodology, and a mixed-methods approach combining qualitative and quantitative research methods. The research was conducted through the following steps:

1. Study 1: Purposed to study perceptions of behavior change in five domains and of digital companions as social actors and collaborators promoting health (40 participated, aged 29–60, see Section 3.1). Based on the results, a framework for designing agent roles and behavior was developed, and a set of argument-based dialogue scenarios were built;
2. Study 2: Extended Study 1 to explore readiness for change in relation to agent roles and behaviors, and perceptions of agent behavior based on the framework (82 participated, aged 29–60). Based on the results a prototype was further developed containing adjusted argument-based dialogue scenarios and a method for tailoring the agent’s behavior and roles; and
3. Study 3: Purposed to evaluate the results from studies 1 and 2 in a formative user study of the prototype involving nine experts (public health, nutrition, epidemiology, nursing, and ethnology): The framework, adaptation methods and argument-based dialogues were introduced, evaluated, and further developed.

For data collection in study 1, a questionnaire was developed and applied in English, which was composed based on a set of baseline assessment questions translated from Swedish, drawn from the prototype applications developed as a part of the research project for behavior change addressing:

- General motives for an activity as value directions: questions about the importance, capability, and satisfaction;
- Areas of activities targeted for behavior change: physical activity, stress, alcohol consumption, and tobacco use; and
- Roles of a digital agent in relation to supporting the change of behavior toward healthier habits.

The data collection in study 2 was also done through a questionnaire, which was again conducted in English, which contained a subset of questionnaire 1, limited to only the domains, physical activity and stress. Questionnaire 2 included, in addition, a set of nine dialogue scenarios between a digital agent and two different tentative users. For each of the dialogues, the participant rated the agent’s behavior, and what role or roles they thought it was enacted in the scenario.¹

The data collected using the questionnaires were analyzed quantitatively to find patterns of preferences among roles and behaviors, and qualitatively using thematic analysis for finding themes among open-ended questions regarding activities/goals, roles, and motivations for the agent’s preferred behaviors.

The qualitative and formative user study (study 3) was conducted as a part of a participatory design process of the digital coach application for promoting behavior change, divided into three occasions. Study 3 was conducted in Swedish using the Swedish user interface of the STAR-C application. For the sake of readability of the article, terms from the study have been translated into English. Ten domain experts were invited to participate, and nine participated in total.

Four participated in the initial individual session in which they used the prototype, containing five adapted dialogue scenarios in addition to the baseline questions, functionality allowing them to set short-term and long-term goals with related arguments and motives, and the set of questions for adapting the coach’s role and behavior. These questions were revised based on the results from the questionnaire study. The participants were interviewed and observed while using the prototype.

A workshop was organized as the second session, where eight domain experts including the four who participated in individual sessions, participated. They were divided into pairs, where the first four participants were paired with each other to start on the same level of knowledge about the system. They were given the task to select activities as goals for behavior change, along with the motives (arguments) for why they want to change, then setting their preferred role or roles and behaviors of the agent. After this, they conducted five dialogues (same as in the individual session). The pairs were instructed to discuss and reflect on the things they experienced and provided examples of how the dialogues ideally would unfold based on their expertise in supporting behavior change. After the sessions in pairs, aspects were discussed with all eight participants. The participants were asked to take notes during the session and were partially observed.

¹ The two questionnaires are found in the [Supplementary material](#).

The results of the second session were used for further modifying dialogues implemented in the dialogue demonstrator, and the new versions were evaluated at a third session in a group with seven domain experts participating, including a ninth expert who had not participated in the earlier sessions. The results were also used for further developing the architecture and the generation of argument-based micro-dialogues.

3.1. Participants

A total of 40 anonymous participants located in Scandinavia were recruited in study 1 through the Prolific service, and 82 participants in study 2, and 122 participants in total. There was an even gender distribution (58 women, 61 men, and three other) among the participants. The participants' age range was between 29 and 60 years (for age distribution, see Table 1). The age range was chosen based on the most prevalent in stress rehabilitation clinics and the age groups participating in the VIP.

Study 3 was conducted as a part of the participatory design process employed in the research program STAR-C, and engaged nine participants (three women and six men) who had been contributing to earlier versions of the prototype in three different sessions (four participated in session 1, eight in session 2, and seven in session 3). The participants had a broad range of expertise, including epidemiology, public health, nutrition, nursing, social work, and ethnology.

3.2. Role and behavior of the digital agent

We defined and exemplified four roles that the participants could relate to and choose from in studies 1 and 2. They could also suggest other roles if the roles proposed did not fit their needs. The participants were asked what role or roles they envisioned digital support could take on among the following:

1. *An assistant* that keeps track of your information and reminds you about what you want to be reminded about;
2. *A coach*, similar to a personal trainer who challenges and encourages you to do things;
3. A kind of *health expert*, which informs about the current state of knowledge and gives advice; and
4. *A companion* that is more like a friend, keeping you company and is on your side.

The participants were then asked to provide a scenario and motivate the previous answers.

TABLE 1 Age of participants.

Age	Study 1	Study 2	Study 3
Below 30	1 (2%)	1 (1%)	0
30–39	21 (53%)	60 (74%)	0
40–49	12 (30%)	11 (14%)	4
50+	6 (15%)	9 (11%)	5
Summary	40	81*	9

*One participant who provided erroneous information about age was excluded in the overview.

In study 2, the participants could also assign behaviors to their preferred type of coaching agent along the following: *how brief*, *how fact-based*, *how challenging*, *how emphatic*, and *how friendly*. The participants could select a value on a four-item scale ranging between *very* and *not particularly* in the first three, and the scale had a middle value for the last two labeled *neutral*. This way, a participant could choose a value corresponding to “un-friendly” if they found the agent behaving this way.

After the participants had provided their own wishes for a digital coach, they applied these roles and behaviors to assess the agent's behavior in the argument-based dialogue scenarios.

3.3. Framework for adapting the agent's behavior

A framework for adapting the agent's behavior was developed based on study 1 and was further refined based on the subsequent studies. Statements describing the agent's preferred behavior and roles were thematically analyzed and clustered into themes of behaviors and roles. As there were differences among the 40 participants, which seemed to relate to which stage they are in the process of changing behavior, more specific questions to categorize a participant into one of the stages of the transtheoretical model of behavior change (TTM) (Prochaska et al., 2015) were added in study 2.

TTM was first introduced by Prochaska and Di Clemente in the late 1970s and was constructed by six stages of behavior change: *Precontemplation*, *Contemplation*, *Preparation*, *Action*, *Maintenance*, and *Termination*. Persons in the *Precontemplation* stage do not intend on taking action, in our case within the next 3 months. When it comes to people in the *Contemplation* phase of the stages of behavior change, they are ambivalent toward changing their behavior. The *Preparation* stage is where some people are trying to change and have intentions of changing within the next month. *Action* is when the person has been practicing the new behavior for a short period of time, usually between 3 and 6 months. People in the *Maintenance* stage are already motivated and committed to the behavior change and have been doing the activity for longer than 6 months.

The framework is outlined in Figure 1. Some comments provided by participants in the first study are exemplified, along with roles, and stages of change based on two complementary dimensions: One is the extent of empathy and friendliness, and the second is the extent of emotional challenge. This framework was used for designing the nine dialogue scenarios in study 2 and the five scenarios in study 3. An analysis of the data collected in study 2 was conducted for exploring to what extent the choice of agent behavior and role related to what stage of change the participant was in. Furthermore, the roles were further evaluated qualitatively from a user experience perspective in study 3. In the following section, the dialogue scenarios are presented.

3.3.1. Dialogue scenarios in studies 2 and 3

The dialogue scenarios were designed based on the behaviors of preferred coaching agents described by the participants in study 1. The dialogues were engineered with the intent of illustrating how brief, facts-based, challenging, or empathic/friendly an agent can be during the scenarios. Dialogue scenarios containing two characters,

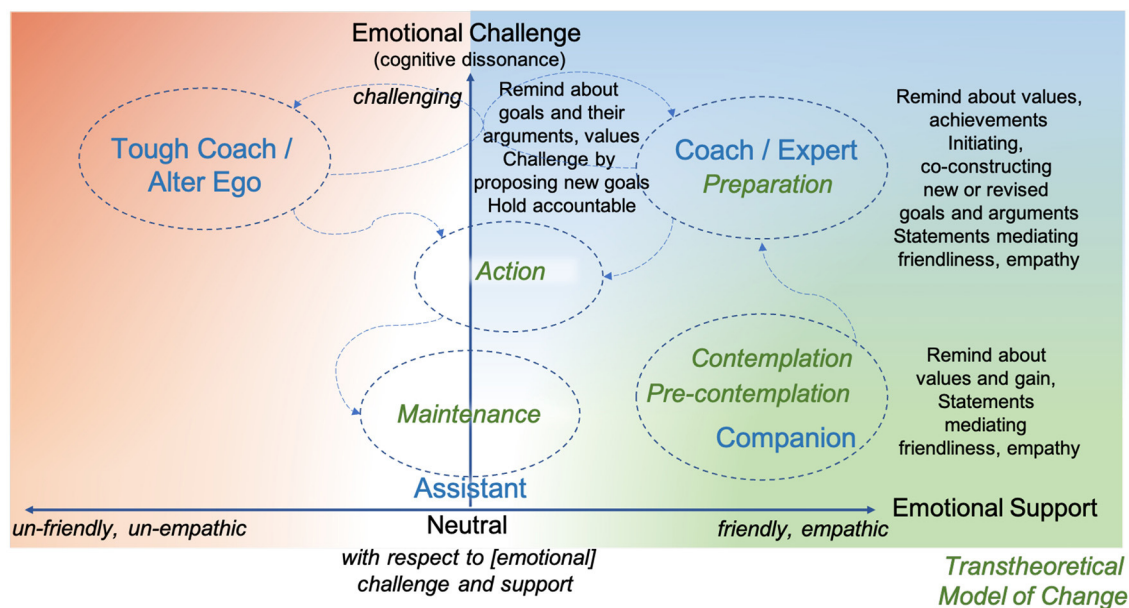


FIGURE 1

Framework for mapping behaviors of a digital agent along two dimensions: *emotional support* (horizontal) and *emotional challenge* (vertical); to *roles* (blue) and to *stages* in the transtheoretical model for behavior change (TTM) (green), mapped out based on the results of study 1. The framework was applied and evaluated in studies 2 and 3. The arrows represent desirable transitions between TTM stages ending in a stable state of maintenance; there are also potential transitions between roles with a switch to the tough coach and back. The color scheme is inspired by compassion-focused therapy, which uses *green* to represent rest and comfort (soothing), *blue* for energy and action (drive), and *red* as a state of conflict (threat) (Gilbert, 2009). Desired actions provided by participants are exemplified.

Jim and Kim, during different parts of the day/days were authored based on the tentative answers the characters could provide on the baseline questions of the behavior change application, also embedded in studies 1 and 2. The two characters differed, where Jim was more focused on increasing physical activity, and Kim was more focused on managing stress (Figure 2). The nine dialogue scenarios contained between two and 13 statements, 74 in total, with an average of eight statements.

In Example 1b, given in Figure 2, a deliberative dialogue is taking place between the digital agent and user Jim, mutually trying to reach a solution through finding common action. By holding Jim accountable through reminding later in the day and being not completely neutral with respect to emotional challenge and support, the agent portrays characteristics of a brief, superficially friendly, mainly challenging coach (1b in Table 2).

Different types of argumentation dialogue were assigned to different scenarios while maintaining uniformity with the framework in Figure 1. The dialogue types used in the scenarios are *Information-seeking*, *Deliberative* (deciding about what to do), and *Persuasive* (changing the attitude or behavior of the other agent), as defined by Walton and Krabbe (1995). We complemented these types with a type suitable for the application in focus, which we call *Supportive* to elicit arguments primarily aimed at providing emotional support embedding empathy.

An outline of the characters and types of dialogues with respect to the scenarios can be seen in Table 2. As can be seen, most dialogues consist of elements from different dialogue types.

The five characters applied in the five micro-dialogue scenarios in study 3 were defined based on the model in Figure 1 and on other results of study 2. The characters were named using gender-neutral

terms—we chose the Spanish words for numbers (Table 6)—and their characters are illustrated in Figure 3.

3.4. STAR-C prototype applied in study 3

The prototype applied in study 3 is a mobile application covering the behavior change domains' *physical activity*, *stress management*, *nutrition*, and *alcohol and tobacco consumption*. The application contains the following:

- A baseline assessment based on the VIP health assessment consisting of a set of questions, of which a subset was used in studies 1 and 2.
- Goal setting by defining activities to be performed within the coming days/week(s), related to behavior change domains, partly also embedded in studies 1 and 2.
- Setting the roles and behaviors of the digital agent, also embedded in study 2.
- Dialogue demonstrator for evaluating five digital agent characters for the purpose of study 3.

The development of content and structure of the application is done using the content management system ACKTUS, which is a platform for knowledge engineering and design (Lindgren and Yan, 2015). ACKTUS contains a core ontology stored in a graph database (RDF4J²) based on the World Health Organization's International

² <https://rdf4j.org/>

<p>Jim wants to increase physical activity to improve health, manage a pain condition and reduce some body weight. He also wants to reduce stress. He has a goal to take 20-minutes walks with his dog four times every day.</p> <p>The main barriers that he faces are that he finds himself too lazy, having too little time and in addition, he does not think exercise is fun.</p> <p>At lunchtime the digital agent initiates a conversation with Jim that unfolds in the following way:</p>		<p>Kim wants to decrease stress caused by work related responsibilities. The importance to do this he rated very high, at the same time his satisfaction was rated very low.</p> <p>He also thinks physical activity is important but isn't very prepared or confident in achieving this change.</p> <p>He has a plan to play games with his partner or friends as a recovery activity to alleviate stress three times per week.</p> <p>The main barriers he has is finding time, and in particular, mutual free time with his partner or friends, since work takes extra time.</p> <p>Kim initiates a conversation with the digital agent after coming home from work, and the following conversation takes place:</p>	
Digital Agent	Hi, lunch time and time to walk with the dog!	Digital Agent	Hey Kim, What's on your mind?
Jim	I don't have the time!	Kim	I feel overwhelmed with the stress I have at work.
Digital Agent	You want to exercise or do other physical activity because you want to reduce pain, remember?!	Digital Agent	Stress can affect your memory and affect your ability to work in the long run, which you probably already are aware of. How much on a scale from 0-10?
Jim	Work is more important now.	Kim	8
Digital Agent	The dog would be happy!	Digital Agent	This is high, and as high as you have rated the past three weeks. To have high levels of stress over a long period of time is not good, it is important to break this and prioritize time to recover. Is it something on your mind that you should/can do/take care of right now?
Jim	I'll take a longer walk tonight.	Kim	No, it is upcoming things next week I am thinking of.
Digital Agent	Okej, it's a plan, I will remind you!	Digital Agent	Then it is important to take the opportunity now to rest and recover! How about playing games with partner or friends again? It usually helps you to de-stress, you know, and it is part of your plans to do this 3 times per week. So far, you have set this up once a week the past month. It would be really good if you can do this more often as you planned to do.
		Kim	That's a good idea. If not tonight, we can set up a game night at the weekend.
		Digital Agent	How about as soon as possible? Could be a short session. Let me know what they say!
		Kim	Okej, I will call them.
		Digital Agent	Great!

FIGURE 2

Example of two scenarios: Jim having a short dialogue at lunchtime (1b), and Kim initiating a dialogue at dinner time (2a).

Classification of Function, Ability, and Health (ICF),³ which is extended with specific and relevant sub-concepts in the class *Personal Factors* and in the class *Activity and Participation*. Other classes are *Body Function and Structures*, and *Environment*, containing social relations and support. ICF is complemented with the class *Diseases and Syndromes* for capturing medical and health conditions.

The ACKTUS ontology also embeds a modified version of the AIF, developed for the purpose to exchange arguments over the web (Chesñevar et al., 2006). An argument (scheme node) is a composite structure consisting of a set of premise nodes (information nodes or *i-nodes*) connected to a conclusion node (*i-node*) in the graphical database. A premise node relates to information obtained from the user when using the application in the baseline assessment, when setting goals, assessing progress, or in dialogues with the agent. An *i-node* in ACKTUS is typically linked to a *value*, which can be any that the content modeler decides. Examples of key values in this application supporting behavior change are importance, satisfaction, how fun, how confident, and how prepared a user is to change

behavior. Furthermore, the node is also linked to a *concept*, e.g., an *activity* (process) in the Activity and Participation class (e.g., physical activity), or to *objects*, such as body functions and structures, diseases, or relationships. The concept informs about what topic is at focus in a dialogue. In a *deliberation* dialogue, the topic is related to the class Activity and Participation, while in an *inquiry* dialogue, which has the purpose to build new knowledge it relates to a class of objects. Consequently, a conclusion of an argument can be related to an activity (about what to do), an object (about what we know), or an advice.

In ACKTUS, the conclusion node can be of three types: (i) a *decision*, such as in the case of a medical diagnosis, with a value; (ii) an *activity*, in the form of an assessment protocol for what to do next (e.g., a set of follow-up questions); or (iii) an *advice*, or *piece of information*. These correspond to the argumentation dialogue types mentioned earlier (i) inquiry dialogue; (ii) information seeking or deliberation dialogue; and (iii) persuasive or supportive dialogue. Each composition of premise nodes and a conclusion is associated to an argumentation scheme, which is also modeled in ACKTUS. At the time of conducting study 3, all arguments were associated with the scheme *argument from expert opinion* since the application at

³ <https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health>

TABLE 2 Scenarios.

Persona	Scenario	Time	Character	Dialogue type
Jim				
	1a	Morning	Neutral assistant	Deliberation
	1b	Lunch	Brief, superficially friendly, mainly challenging coach	Persuasion and deliberation
	1c	Next morning	Friendly, challenging factual expert	Persuasion and deliberation
	1d	Lunch	Non-challenging, brief, friendly and empathic companion	Information-seeking, supportive and deliberation
	1e	Next morning	Non-brief, challenging expert	Persuasion and deliberation
Kim				
	2a	Dinner	Non-brief, challenging expert	Persuasion and deliberation
	2b	Next morning	Factual, Friendly and empathic companion	Information-seeking, supportive, and deliberation
	2c	Dinner	Factual neutral assistant wrt emotional support	Information-seeking, supportive, and persuasion
	2d	Next morning	Brief coach, challenging by goal-reminders, and holding accountable	Information-seeking, deliberation

that point contained only knowledge engineered by medical domain experts.

The dialogue demonstrator contained a short description of the Jim scenario, on which the five characters' dialogues were built. The dialogues were modeled using ACKTUS. In the initial step, the user was given three answering alternatives: *positive*, *neutral*, and *negative* for each statement provided by the agent. The next statement posed by the agent depended on the response made by the user. The participants were instructed to select the response based on how they experienced the statement, e.g., liked the statement, or agreed with the statement, or not. Focus was on their experiences and on exploring different ways to respond to the agent's behavior, role, and attitude. Based on the participating domain experts suggestions, the dialogues were modified to encompassing different kinds of responses, which were evaluated by domain experts in a third session.

4. Results

The results are organized as follows. In the following section, the readiness levels based on TTM assessed in study 2 are summarized, and the participants' views on motives and barriers for changing behavior. The participants' own expectations of a digital coach or companion in terms of roles and behaviors, and their relation to TTM levels summarized in Section 4.2. The participants' perceptions of the exemplified agents taking on roles and behaviors in the scenarios are presented in section 4.3.

The results from the three studies feed into ongoing work on further developing the architecture and argumentation process for generating person-tailored argument-based micro-dialogues. The argumentation process is introduced and exemplified in section 4.4.

4.1. Participants' view on motives for changing behavior related to physical activity and stress

Among the 82 participants in study 2, 19% had always been physically active, and 24% had always been able to manage their stress levels. We consider these being in the maintenance stage of

the TTM model (Table 3). For physical activity, a vast majority (75%) is considering changing their behavior within the coming month or within 3 months. A difference is seen in changing behavior to reduce stress, where 30% is planning to make a change. While 23% have a good balance for managing stress, and another 20% has no plans for change coming 3 months, as many as 23% expects an increase in levels of stress (Table 3).

The participants' motives relating to a value direction serve as arguments on the needs level of human activity, which is connected to an activity set as goal in the studies (Table 4). The motives were crossing over the two domains for behavior change, such as physical activity was motivated for some as recovery activity from stress which was noticeable in how the participants defined other reasons than those suggested. Furthermore, arguments motivating the choice of value direction, as well as barriers, are captured (Table 5).

A low proportion of the participants chose social motivators for their chosen baby-step activity to increase physical activity, social motivators being others' expectations, keeping up with society, and nurturing relationships with friends and family (Table 4). A similar pattern is seen for the baby-step activity to reduce stress, where nurturing relationships with immediate family motivated 22% of the participants. An interesting observation is that the participants seem to have chosen baby-step activities that they find being fun and/or entertaining to a large extent for mitigating stress (63%).

When analyzing the motivators based on gender for their chosen baby-step activity, the answers given were similar in the amount of male and female participants in physical activity as well as for stress. The most apparent reasons for doing their chosen physical activity were physical wellbeing (79% of women and 85% of men), emotional wellbeing (59% of women and 69% of men), and it gives energy (62% of women and 52% of men).

4.2. Expectations related to the digital coach's role and behaviors in dialogues

The participants in studies 1 and 2 were asked what role or roles they envisioned digital support could take on among the following (proportion of participants in parentheses) (Table 3): (i) an *assistant*

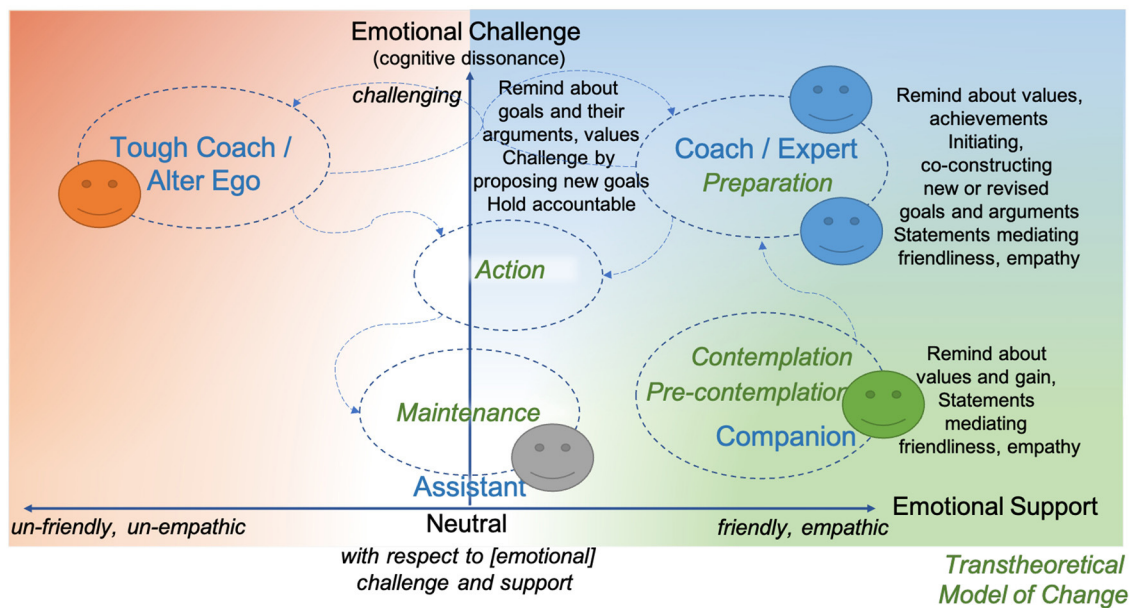


FIGURE 3

A total of five characters interpreted in the framework for mapping behaviors of a digital agent, Uno is colored green, Dos is gray, Tres and Cuatro are blue, and Cinco is orange. The arrows represent desirable transitions between TTM stages ending in a stable state of maintenance, there are also potential transitions between roles with a switch to the tough coach and back. The color scheme of the agents follows the colors of the compassion-focused therapy as in Figure 1, with the complementary color gray for the neutral assistant.

TABLE 3 Study 2 participants' stage of change (TTM), related to what role they preferred, and summary of all 122 participants' choices of roles.

TTM Stage	Number/Stage	Assistant	Coach	Expert	Companion
Physical activity <i>n</i> = 82					
Precontemplation					
No plans for coming 3 months	4 (4.9%)	25%	0%	50%	0%
Contemplation					
Plan to change within 3 months	32 (39%)	69%	50%	34%	28%
Preparation					
Plan to change within 4 weeks	30 (36.6%)	67%	63%	53%	20%
Action					
Have started to change	N/A	N/A	N/A	N/A	N/A
Maintenance					
Change since more than 6 months	15 (18.3%)	40%	60%	27%	20%
Stress <i>n</i> = 82					
Precontemplation	17 (20.1%)	50%	31%	31%	19%
Contemplation	9 (11%)	44%	67%	22%	22%
Preparation	21 (25.6%)	67%	67%	33%	29%
Action	N/A	N/A	N/A	N/A	N/A
Maintenance	19 (23.2%)	63%	58%	58%	21%
Termination-risk for relapse	19 (23.2%)	69%	56%	44%	25%
All in study 1 and 2	<i>n</i> = 122	73 (61%)	68 (57%)	45 (39%)	28 (23%)

(61%), (ii) a coach (57%), (iii) a kind of health expert (39%), and (iv) a companion (23%), and two participants preferred it to not having a role at all.

The participants were also asked to provide a scenario and motivate the previous answers. An overview of the themes that emerged is shown in Figure 4. Two major purposes emerged that

TABLE 4 Motivations in terms of value directions (*vd*) for the participant's chosen baby-step activity to increase physical activity or decrease stress.

I do the activity because	Physical activity <i>n</i> = 122	Stress <i>n</i> = 122
<i>vd</i> ₁ . It gives energy	68 (56%)	50 (41%)
<i>vd</i> ₂ . It's fun, entertaining	47 (38%)	77 (63%)
<i>vd</i> ₃ . Rest and recover	29 (34%)	91 (75%)
<i>vd</i> ₄ . Others' expectations	10 (8.2%)	3 (2.5%)
<i>vd</i> ₅ . Obligations	15 (12.3%)	3 (2.5%)
<i>vd</i> ₆ . Improve physical wellbeing	101 (83%)	44 (36%)
<i>vd</i> ₇ . Nurture relationships with immediate family	10 (8.2%)	27 (22%)
<i>vd</i> ₈ . Nurture relationships with friends and social network	16 (13%)	14 (11.5%)
<i>vd</i> ₉ . Keep up with society	7 (5.7%)	8 (6.5%)
<i>vd</i> ₁₀ . Improve emotional wellbeing	76 (62%)	64 (52%)
<i>vd</i> ₁₁ . Other: improve appearance, feel more comfortable, escapism, investment in physical and mental health	6 (5%)	4 (3.3%)

related to either the digital companion more as a neutral assistant or health expert, or as an engaging coach or companion.

The digital assistant would help track and summarize accomplishments and failures and provide reminders for the person to adhere to their goals. This was also perceived as task for a digital coach. The digital assistant was viewed mostly in comparison to a fitness tracker that is available through smartwatches and mobile applications in the market today. The three main themes that appear under the digital assistant umbrella are *simple informer*, *reminder companion*, and *fitness tracker*. Uses for the digital assistant in the views of the participants were activities related to such as tracking of sleep and calories but also informing and reminding of the to-dos. Although few similar expectations were summarized under the digital coach and the digital assistant roles, variance of participants' expectations between the two roles is clearly apparent. The digital coach themes were *challenging coach*, *authority figure*, *professional trainer*, and *goal-setter*, and it was expected to hold the participant accountable and keep its user on track toward his/her goal through challenge and encouragement. Some participants also wanted the digital coach to embed steps on how to conduct certain tailored physical activities depending on the user's situation.

As for the digital health expert, it would provide personally relevant information and new knowledge, including fearful facts about the consequences if changes are not made to improve health. The main themes that appear in a digital health expert are *advisor* and *monitor* of health status and diagnostics. The *advisor* health expert, in views of the participants, would apprise and recommend for, for instance, preemptive actions against mood dips and adapt to the needs of the user's status related to injury and rest time.

The other categories of purposes related to personal and emotional support are then delivered by a digital coach or companion. Purposes include keeping company, encouragement, motivation, giving inspiration, maintaining reasonable expectations, maintaining discipline, challenge, holding one accountable, telling

what to do, and pushing to do activities. Moreover, it could add some fun.

The digital companion role mostly encompassed emotional support and company. The companion was envisioned to be a relief from stressful events and a replacement for human partners in the case of them not being available. The participants also expected the digital companion to be adaptable and unbossy while maintaining its pushy-friendly behaviors.

Furthermore, the relationship between the stages of the TTM and preferred roles (*assistant*, *health expert*, *coach*, and *companion*) and behaviors (*how brief*, *how fact-based*, *how challenging*, *how emphatic*, and *how friendly*) was explored. This was done to see if the preference for a certain type of behavior or role was dependent on the stages of change (Table 3).

A combination of roles was selected by 56%. The most frequently selected role was assistant (61%) and coach (57%), the expert role was selected by 39%, and the least frequently selected was the companion (23%). The assistant role was less preferred by people in the contemplation stage for managing stress, and people in the maintenance stage for physical activity, compared to how often the role was selected by people in other stages. The companion role seemed to be slightly more interesting to people in the contemplation stage for physical activity, and in the preparation stage for managing stress than compared to people in other stages. Moreover, people rating high importance to change behavior to decrease stress preferred a digital companion over other roles.

Figure 5 shows how the preference for empathetic and challenging behavior is distributed over the stages of change. Approximately 10% across the stages wished the agent to be very empathetic, while between 40 and 60% wished it to be not particularly empathetic (Figure 5). The rest desired a neutral digital agent, with respect to empathy. About half of the participants wanted the agent to be challenging to a different extent, half to not be particularly challenging. A difference was seen between physical activity and stress, in which participants who wanted the agent to be challenging leaned more toward preferring the agent to be more challenging when supporting behaviors relating to stress than physical activity.

4.3. Participants' perceptions of the agents' behaviors and roles

The results of study 2 showed that the participants, in some cases, perceived the agent to express more empathy and friendliness than what they were designed to express, which was the main discrepancy in the cases, the participants had a different perspective on characters and roles (characters in scenarios 1e, 2a, and 2c in Table 6). Due to this, the subsequent characters in study 3 were designed to express more clearly friendliness/empathy, neutrality, and absence thereof ("non-friendliness/non-empathy"), respectively.

4.3.1. The participating experts' views

The participants in study 3 reflected on the roles and behaviors of the digital agent in the context of promoting health, while using the prototype application. An overview of their perception of the five example characters is shown in Table 6. While they agreed on the intended characters, roles, and behaviors, what they liked and did not

TABLE 5 Participants' arguments in favor and against changing behavior to increase physical activity.

I want to exercise or do other physical activity because	n = 122	Type of motivator
m ₁ . I want to improve my health	114 (93%)	Introjected regulation
m ₂ . Research shows that physical activity prevents many diseases	71 (58%)	Introjected regulation
m ₃ . I want to reduce pain	35 (29%)	Introjected regulation
m ₄ . It is relaxing	30 (25%)	Intrinsic motivation
m ₅ . It makes me feel good	81 (66%)	Intrinsic motivation
m ₆ . It gives energy	68 (56%)	Identified regulation
m ₇ . It is a social thing	11 (9%)	Identified regulation
m ₈ . I have to because I sit still all day at work	35 (29%)	Introjected regulation
m ₉ . I have always done it, it is a habit	10 (8,2%)	A-motivation
m ₁₀ . Other: reduce weight (3), kids to be active, reduce stress, improve cognition, mental health, sense of accomplishment, feel stronger, treat physical condition	13 (11%)	Misc
I don't exercise/or do physical activity because		Type of barrier
b ₁ . I have never done it regularly, it is not a habit	44 (36%)	Personal: habitual
b ₂ . I cannot find the time for it	44 (36%)	Personal: organizational
b ₃ . I do not think that it is fun	34 (28%)	Personal: emotional
b ₄ . I have too much pain, or other physical condition that stops me	26 (21%)	Physical
b ₅ . The weather is not good	31 (25%)	Environmental
b ₆ . It is too expensive to do the things I want to do	15 (12,3%)	Socio-economic
b ₇ . I would like to do it with others, who are not available	17 (14%)	Social
b ₈ . Other: depression (2), not enough energy (2), lack of discipline, long distance, fear of falling, others' judgment, laziness, have a baby	16 (13%)	Misc

like varied. Uno was preferred by one who found it to be encouraging and “here and now.” The most preferred character was Tres, the empathic and challenging coach/expert, followed by Cinco, the non-friendly and challenging coach. Those who preferred Cinco found it intriguing, “a little evil,” and fun, compared to the other examples, and as a way to “push.” They found it being good that it is straight to the point and good for the memory to be reminded.

Those who liked Tres the most, also disliked Cinco the most, using words like “terrible,” “not acceptable.” One of the participants who preferred Tres and disliked Cinco motivated this by wanting a digital companion or coach who could provide a basic sense of comfort, safety, and trust, which would not work with Cinco. On the other hand, when the basic foundation of trust and comfort is established, the agent could in some moment turn into the Cinco character to provoke/challenge the participant's attitude: “...then it can be ok with more harsh comments as a kick in the butt.” More comments on that a variation in behavior and a mix of attitudes were preferred, both “soft, compassionate but could be firm.”

General comments concerned the amount of information about health in the statements provided by the digital agent. Shorter, to-the-point statements about health were desired; better to be more briefer than too facts-based and lengthy in arguing why changing behavior is desired. Suggestions of dialogue elements included ending with a question that the person can respond to, which also works as a challenge, something to think about.

Alternative ways for the user to respond to arguments were suggested, partly to make the user reflect and collect the user's

view on the argument, partly to lead the reasoning process forward toward a positive conclusion about what to do. In addition to information-seeking purposes, the following three general responses were identified:

- (i) to state *confirm*, *reject* (potentially moving forward in time), or *undecided* (expressing ambivalence);
- (ii) *confirm*, *reject*, or *undecided* as in previous but also including a *reason* for this among barriers or motivators identified as relevant to the individual (pose a supporting or attacking argument); or
- (iii) to reason about what *emotional support* or *challenge* the individual needs in the current moment (change topic to how to act).

Examples were embedded in new versions of the five dialogue scenarios and discussed at a follow-up session with the experts. While confirming that their perspectives and suggestions were embedded in the new versions, they also highlighted the cultural aspects concerning *how* to express things in dialogue with different people.

4.4. Person-tailored argument-based micro-dialogues

The application STAR-C used in the study is being developed to embed a digital coach, which utilizes value-based argumentation embedding supporting and challenging arguments. When developing

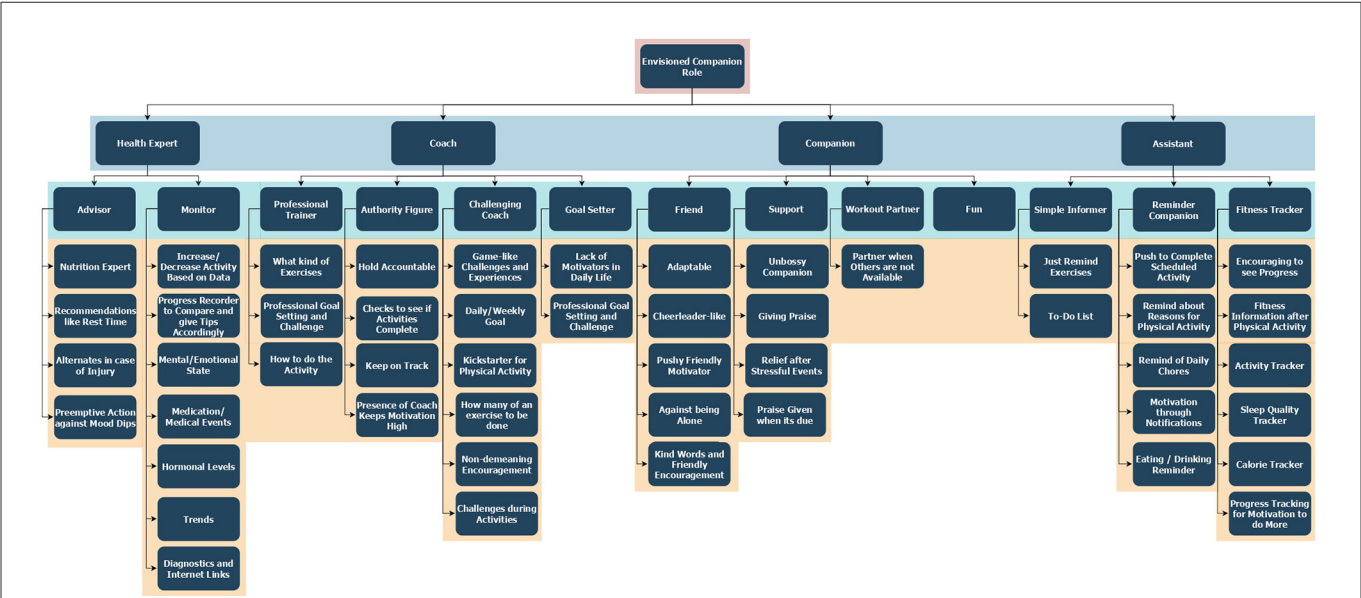


FIGURE 4 Resulting themes based on study 2 participants' views on the agent's behavior and roles. The blue layer outlines the envisioned companion roles participants have described, whereas the turquoise layer describes the sub-roles the companion can play. The orange layer describes reasons for choosing a sub-role or actions participants would want a companion to execute.

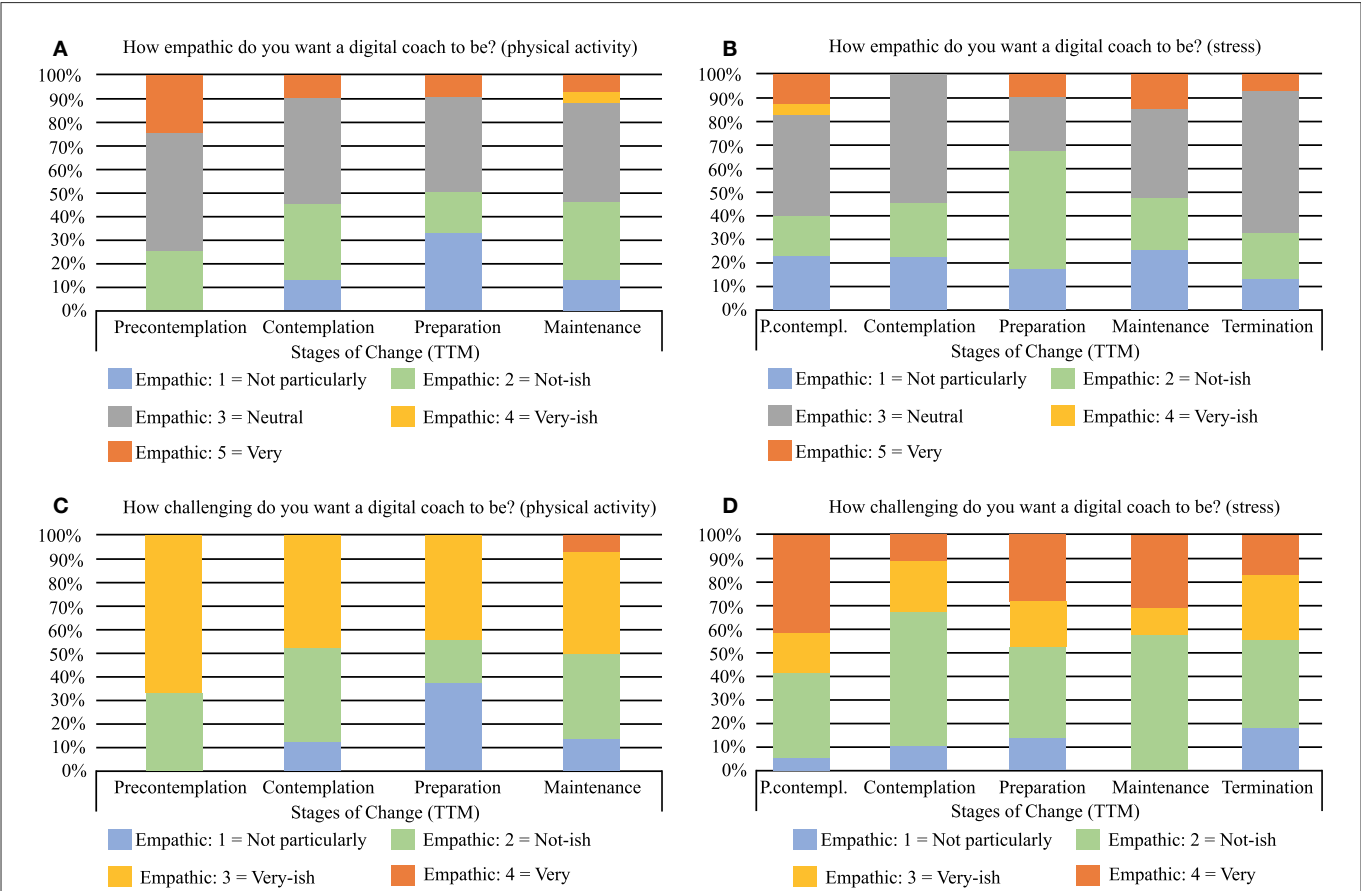


FIGURE 5 Preferred behaviors of a digital coach, for the different stages of the TTM, for physical activity (A, C) and stress (B, D). To be noted: For physical activity, four were in the precontemplation stage, and only one participant was in the termination stage and was therefore omitted in the overview, see Table 3.

TABLE 6 Comparison of defined and perceived character and roles.

Char. (Scen.)	Defined character	Perceived character	Comment
Study 2			
(1a)	Neutral assistant	Also bit friendly/empathic	Agreement
(1b)	Bit friendly, challenging coach	Same	Agreement
(1c)	Friendly, challenging expert	Also coach	Agreement
(1d)	Non-challenging, friendly/empathic companion	Same	Agreement
(1e)	Challenging expert	Neutral coach	Difference*
(2a)	Challenging expert	More coach/companion	Difference*
(2b)	Non-challenging, friendly/empathic companion	Same	Agreement
(2c)	Neutral assistant	More companion	Difference*
(2d)	Challenging coach	Also friendly/empathic	Agreement
Study 3			
Uno	Friendly/empathic companion	Empathic, caring, too friendly	Agreement
Dos	Brief neutral assistant	Less empathic, focus numbers	Agreement
Tres	Friendly/empathic, challenging coach/expert	Comforting, safe	Agreement
Cuatro	Friendly/empathic companion/coach/expert	More rehab	Agreement
Cinco	Challenging, non-friendly/empathic coach	Little evil, fun; horrific	Agreement

the STAR-C coach module further based on the results of the studies presented in this article, we explore how argumentation schemes can be utilized. The STAR-C mobile application uses the knowledge base embedded in the ACKTUS platform (Lindgren and Yan, 2015), as introduced in Section 3.4. The user's information that is collected at baseline and in daily use will be used by the system for tailoring short dialogues (micro-dialogues) to the individual. In this section, a high-level description of the construction, evaluation, and the application of arguments in dialogues with a user is presented. Furthermore, the findings presented from studies are applied in an example case based on one of the participants. The purpose is to exemplify the adaptation of roles and behaviors to the individual's preferences, goals and values, and the argumentation process. Also, the different types of responses in the dialogues are exemplified.

4.4.1. Representing generic knowledge and knowledge about the user

The following is an example of how an argumentation between a digital agent and one of the participant from our study, Jane (alias), could play out based on Jane's value directions, actions, motives, and preferences regarding the digital agent.

Jane wants to increase physical activity to improve health, which she rates most important, and lose weight. She also wants to reduce stress, which she rates as very important. She has a goal of walking her dog for 30 min per day and has stated to the digital companion that walking her dog is the best method for dealing with stress, as recovery activity, and that she has to do it. Therefore, Jane wants her digital agent to be a companion with some empathy, but also a "Tough Coach/Alter Ego" to challenge her and be pushy at times to support her to reach her goal.

The main barriers Jane faces is that she lacks energy, thinks exercising is not fun and the weather where she lives is usually bad. Moreover, she often does not have the time.

At baseline, our example user Jane had assessed what behaviors (bh_i) she prioritized to change and selected increasing physical activity (bh_1) and activities to decrease stress (bh_2). For each of these, she assessed *how important*, *how prepared* she is to make a change, *how confident* she is to succeed, and *how satisfied* she is with the current situation. We will, in the following example, apply only the *importance* value and assume she is in the *preparation* stage of TTM, aiming to take action within the coming weeks. At baseline, she had also assessed what is *motivating* her to change behavior relating to physical activity: m_1 (improve health), m_{10a} (reduce stress), and m_{10b} (reduce weight); and *barriers* (i.e., counter arguments) for changing behavior: b_2 , b_5 , and b_8 (Table 5).

At run-time, when defining an activity meeting a short-term goal, the user selects which behavior the activity aims to change (e.g., too little physical activity and/or stress), what they aim to do (Jane in our example is walking her dog 30 min four times per day) how *important* (value) the activity is and how *fun* she expects it to be (value), and with whom they would like to do the activity with (in our example, Jane selected her pet for her walk with the dog). Furthermore, motives related to value directions (vd_i) for taking a walk with the dog are captured (vd_3 , vd_4 in Table 4), as well as the *social* parameter with whom or what the activity is planned to be done, which in our example, also tells who may be disappointed if this activity will not be done. The *goal* is set to do the activity for 30 min four times per day.

In addition to person-specific knowledge, the agent has general knowledge applicable in Jane's case, which it can retrieve from its knowledge base (Figure 8). General knowledge is formulated

as *generic arguments* (*ga*). Each argument is associated with an *argumentation scheme* (*as*). Two schemes defined by Walton et al. (2008) were applied: *argument from expert opinion* (*as*₁) and *argument from position to know* (*as*₂), as exemplified as follows:

- ga*₁ Physical activity increases energy levels (*argument from expert opinion*).
- ga*₂ Recovery activities are necessary to decrease stress levels (*argument from expert opinion*).
- ga*₃ Humans and other animals become happy when socializing and unhappy when opportunities are missed socializing (*argument from position to know*).
- ga*₄ A happy state increases energy and decreases stress levels (*argument from position to know*).
- ga*₅ Increased energy levels make one a better worker (*argument from position to know*).

The first two statements are asserted to be true by experts in the domain of stress management; subject domain is, in this case, psychology. The following three are generic assumptions from positions to know, which can be seen as examples of statements by a person sharing their own experiences with others. Consequently, arguments associated with the different argumentation schemes are ranked differently reliable for instance, an argument from the expert opinion grounded in relevant clinical experiences can be considered stronger than an argument from position to know (Lindgren and Yan, 2015). However, to an individual, the argument that the dog will be happy may be a more personally relevant and, therefore, stronger argument than one based on expert opinion.

The studies presented, in this article, explored argumentation from the additional positions providing emotional *support* for the purpose of providing a sense of being on their side and *challenge*, which may increase cognitive dissonance and tension. These purposes are different from the purposes information seeking, inquiry, deliberation, and persuasion dialogues as defined by Walton et al. (2008). Therefore, to encompass argumentation with purposes other than those defined by Walton et al. (2008), two argumentation schemes were defined: *argument from position to support* (*as*₃) (Figure 6) and *argument from position to create tension* (*as*₄) (Figure 7).

A barrier *b* is identified as something preventing the person (*ag*₂) from doing a desired activity and can be viewed as an argument for why a person would not pursue his/her goal *G* (Figure 6). In the situation when the person's argument for not doing the intended activity that would pursue the goal (e.g., being too tired to do physical exercise) is questioned (attacked or undercut) by the digital agent or other (e.g., physical activity gives you energy), the agent complying with the argument from the position to support scheme would take the supporting position and state, for example, the following:

- ga*₆ There are good reasons not to conduct the planned activity targeting the desired goal, so based on the highlighted circumstances; it is better not to do it at this point (*argument from position to support*).

On the other hand, if the agent would instead comply with the argument from position to create tension, knowing that the person wants to be challenged by the agent, then the agent is allowed

(permitted) to create tension evoking some cognitive dissonance or other emotional engagement to overcome the barrier. However, if the person has stated that challenging behavior is not desired, the agent is not permitted to create tension even if the agent assesses this to be the best strategy based on other factors. The following is an example:

- ga*₇ Weather should not prevent people from conducting activities since people are not made of sugar (*argument from position to create tension*).

These argumentation schemes can be used by the agent to adapt its reasoning to a situation, and reason from which position (role and character) the agent takes on expert, coach, companion, and assistant or the challenging alter ego, this is based on a mutual agreement on the social norms to be applied in the dialogue.

4.4.2. Building and using arguments

The following is a brief overview of the process of constructing and applying arguments in a dialogue, as shown in Figure 8. The approach was inspired by Ballnat and Gordon (2010) argumentation process and the *sufficient condition scheme* based on Walton and Krabbe (1995), which was extended by Atkinson et al. (2006) to embed values. The blue arrows in the figure follow the argument to be constructed. The green arrows follow the path to a dialogue with the user.

When the dialogue is activated by the user or the agent, this triggers the *Construct Arguments* module which fetches the relevant goals, values, activities, and arguments connected to the user. The module puts this information into the relevant contextual information fetched from the *Knowledge Base* confirms adherence to rules and guidelines, and construct arguments utilizing the information. After the construction of the arguments, the *Formulate Arguments* module translates the arguments into a culturally adapted format suitable for a dialogue (e.g., language, language suitable for subgroups in society). The arguments are then recorded with the *Record Arguments* module to be sent into the repository for utilization in future dialogues and arguments.

The arguments, after being recorded in the database, are referred to the *Evaluate Arguments* module to be used in dialogue with the user. The evaluated arguments are then dispatched to the *Compute Position* module. The *Compute Position* module takes on the important duty of combining the behavior and role of the coach, depending on the situation of the user (explained in more detail with examples below) but also is the module which sends the supporting argument or counterargument to be displayed to the user for the continuation of the dialogue. There is always the possibility of the user having something that does not allow them to do the activity suggested or reminded about by the digital companion. The *Argument Left to be Made* component in the digital companion ends the dialogue in a proactive manner, as shown in the dialogue with Jim in Figure 2, if that is the case or when there are no more arguments to be made. If there is room to propose additional supportive arguments or counterarguments into the dialogue with the user, the green arrow dialogue loop continues.

To represent the argumentation-based process in a formalized manner, the extension of Walton's (1996) *sufficient condition scheme* laid out by Atkinson et al. (2006) is adopted as the general scheme

Name: Argument from Position to Support

Major Premise:

- Agent *ag1* is in a position to support Agent *ag2* about certain subject domain *P* containing Goal *G*.

Minor Premise:

- If Argument *arg* (in domain *P*) posed by *ag1* or other agent is attacked by *ag2*
- AND attack is in context of Barrier *b*
- AND Agent *ag2* has the desire to receive support from Agent *ag1*
- AND Agent *ag1* has the desire to support Agent *ag2*

Conclusion:

- THEN *ag2* is supported by *ag1* to address *b* in *ag2*'s chosen way

FIGURE 6

Argument from Position to Support.

Name: Argument from Position to Create Tension

Major Premise:

- Agent *ag1* is in a position to challenge Agent *ag2* about a certain subject domain *P* containing Goal *G*

Minor Premise:

- If all attacks are in context of *b*
- AND Agent *ag2* has the desire to (not) be challenged by Agent *ag1*
- AND Agent *ag1* has the desire to challenge Agent *ag2*

Conclusion:

- THEN *ag1* is permitted (not permitted) to create tension to overcome *b*

FIGURE 7

Argument from Position to Create Tension.

for the agent, which can embed arguments from different positions rooted in other argument schemes. Argumentation schemes function as templates for reasoning, in this example, embedding a positive prediction of the effects of performing the activity the user had planned, both on the action and value-direction levels of activity. The scheme in Atkinson et al. (2006) is given as follows:

as₅: In the current circumstances *R*, we should perform action *A*, which will result in new circumstances *S*, which will realize goal *G*, which will promote some value *V*.

Since contextual knowledge, such as domain knowledge, is essential when reasoning about health, we further extended this scheme regarding current circumstances by specifying different categories of circumstances. In our example, the agent has the following information about Jane's situation, interpreted in terms of the argumentation scheme and available relevant knowledge retrieved from the knowledge base. Relevance is determined by the domain of behavior change and which role the agent is taking on based on the user's preferences and stage of change:

R: (Current Circumstances)

- AgentPreferences = (lunch-time is a preferred moment to interact with the agent; empathic, challenging companion);
- Goal = (walk the dog 30 min);
- Motives = (*bh1*: increase physical activity (importance-value: most); *m1*: improve health; *m10a*: reduce stress, *m10b*: reduce

weight; for the chosen activity *vd3*: rest and recover; *vd5*: obliged to walk the dog);

- Barriers = (*b8*: may be lacking energy, *b2*: may be lacking time, *b5*: rainy weather);
- GenericKnowledge = (*ga1* - *ga7*);

A: (Actions) Walk the dog for 30 min

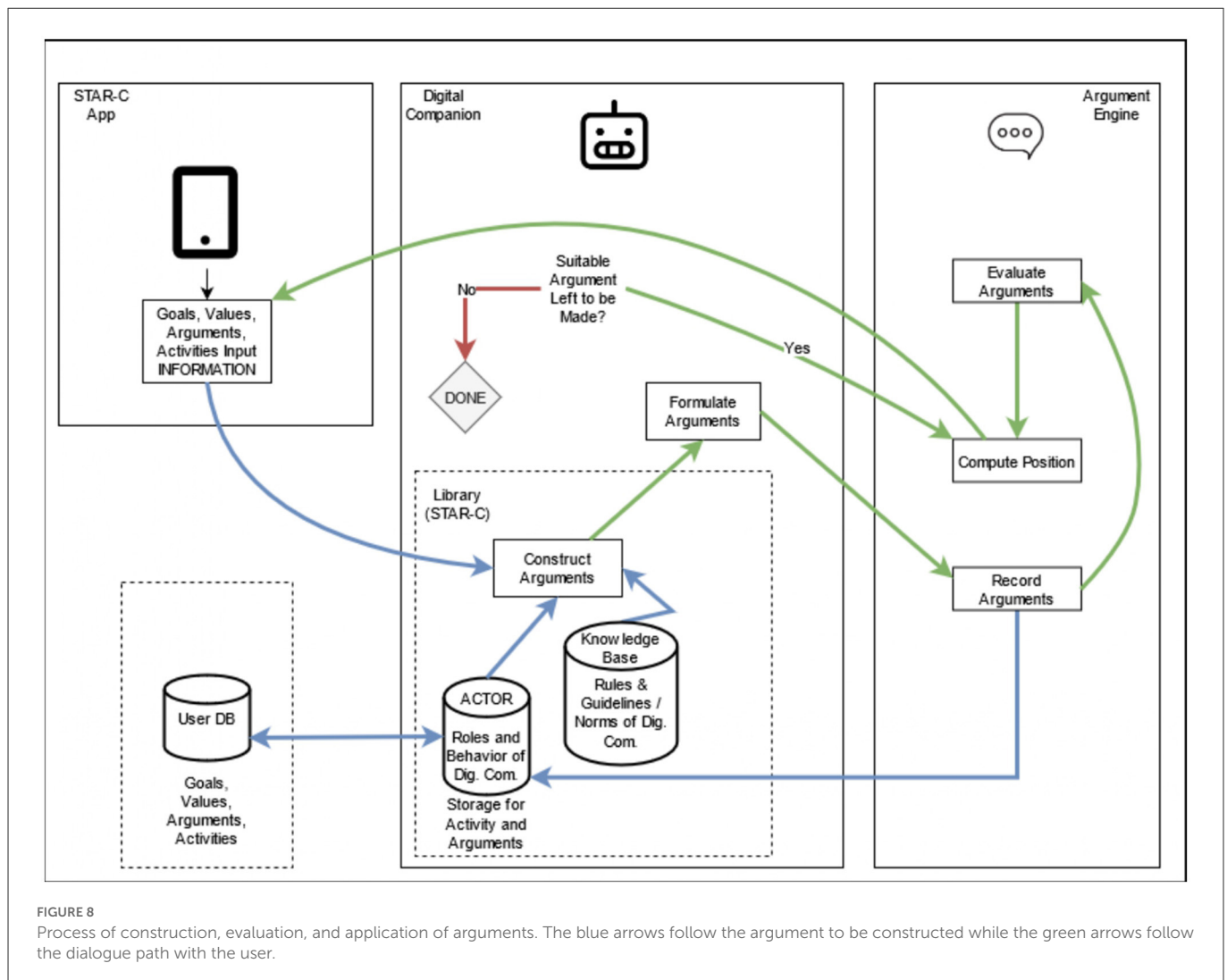
S: (New Circumstances) More energy, Jane and the dog are happy

G: (Achieved Goals) Walked the dog for 30 min

V: (Values) Increased physical activity is most important, reduced stress very important, improved health, reduced weight, and increased energy level.

To continue with our example, at lunch time, the digital companion initiates a dialogue with Jane according to her preferences, with a set of constructed arguments, which are updated during the argument process based on new circumstances provided by the user and with the following set of potential actions, including the activity Jane has specified as the target activity:

1. *Walk Dog 30 min*: The action that follows Jane's plan to increase physical activity,
2. *Walk Dog 15 min*: The action that partially follows Jane's plan to increase physical activity,
3. *Let Dog out in the backyard while having lunch working*: The action that barely follows Jane's plan to increase physical activity but may follow Jane's plan to decrease stress, and



4. *Do Nothing*: The dog is not cared for, so this is not an option due to her obligations.

The dialogue is initiated by the agent, based on the argumentation scheme as_4 ; it poses Argument arg_1 focusing Barrier b_8 , see Figure 9 to see how the dialogue could unfold. One decision point is whether to select a more challenging or more supportive attitude in step 3. Since Jane brings up another barrier (Barrier b_2), the agent follows up in the next step, addressing this barrier.

When Jane brings up yet another barrier, the weather condition (Barrier b_5), the digital agent decides to use the harsher counterarguments, adopting the pushy character as per Jane's choice for persuading her to do it and hold her accountable.

Jane has three alternative responses in the example; in the second alternative, Jane picks up on the potential "loving boot effect" (Blakey and Day, 2012), a stimulation that "kicks" Jane to achieve higher performance, leading the agent to follow-up the walk choose the question about how happy she is afterward. The third alternative is an example when Jane may chose to counteract by changing the topic toward what she needs, rather than what to do (Figure 9).

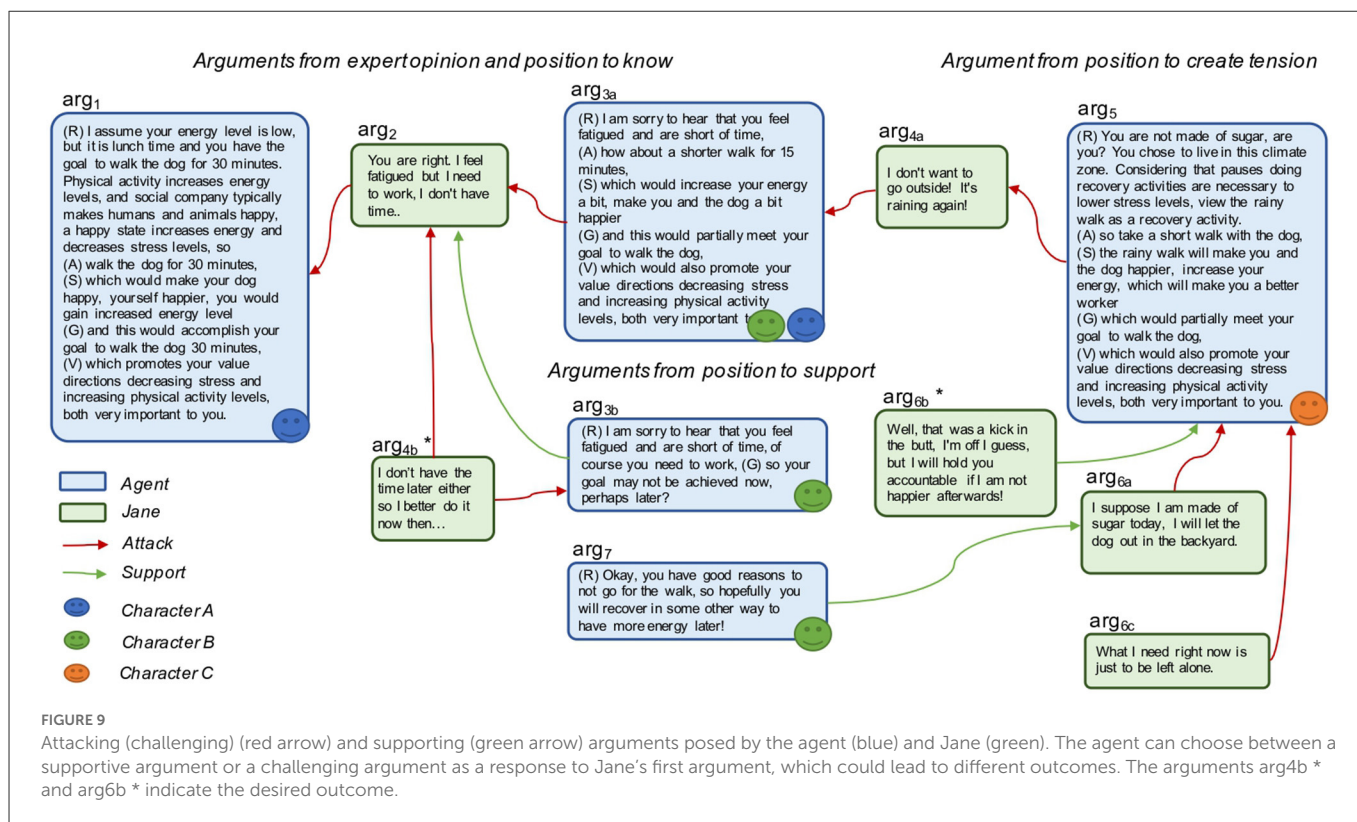
4.4.3. Evaluating and selecting arguments

The agent starts off with selecting a subject domain to target, i.e., topic, based on which assumptions are generated about current circumstances based on the available information and contextual information, such as time of the day.

The order in which the action A is selected relates to the potential options that are available to the agent, the user's selected goals and activities, their assessments of importance and accomplishments so far, and the roles and behaviors preferred by the user.

The agent would rank the set of potential actions based on utility in the value functions (importance and physical activeness in this example since increasing physical activity was ranked highest before reducing stress) and to what extent the action would fulfill the user's short-term goal. The agent would then begin with the option with the highest value, then after evaluating the response from the user and potentially revise the list, go down the list until there is a reason to end the dialogue. Based on the responses of the user and the barriers they have, the agent computes position to be supportive or provoking, along with a re-evaluation of the order of actions.

The subject domain is a factor when evaluating arguments from the agent's perspectives as there are multiple domains in



which the user might want to change behavior. Therefore, varying roles and behaviors might be necessary for certain domains (e.g., a user might be in one TTM stage for increasing physical activity but might be on a different stage when it comes to reducing stress as in this example), while it might not be of the essence in other domains. One strategy the agent can apply is to broaden the subject domain to include more topics (e.g., in our example, also reducing stress) to strengthen the values of conducting an activity when it could serve more than one goal or value direction.

When the user attacks an argument put forth by the agent, the agent must distinguish the barrier that is holding the user from achieving their goal *G*. This is achieved through the ToM the agent has constructed about the user, in combination with the current situation, e.g., weather conditions and time of the day. The counterarguments presented by the user are saved into the repository to be analyzed for future reference and usage in arguments to come.

5. Discussion

The purpose of the research presented, in this article, is to use AI systems to empower individuals to progress in their pursuit of improving health and physical and emotional wellbeing through a change of behavior. This aligns very well with the definition by Nowak et al. (2018) of HCAI as AI that focuses on “collaborating with humans, enhancing their capabilities, and empowering them to better achieve their goals.”

In the notion of *collaboration*, there is a social aspect embedded relating to coordination and agreeing on goals and a division of tasks, typically relating to what roles the actors are enacting. In the studies presented in this article, the digital agent's roles and behaviors as a social actor are explored from the viewpoints of potential users and domain experts, which is discussed in the following section.

Furthermore, when coordinating and agreeing on goals and the division of tasks in an envisioned collaborative journey of the agent teaming up with the user, instruments for the agent to apply are key.

Natural argumentation allowing the user to respond in any way they like would allow the user to express themselves freely and with the language they usually use. However, in this study, structured dialogues are used for the purpose of allowing domain experts to evaluate and verify the agent's behavior, as well as to obtain structured information from the user for feedback and research purposes. The STAR-C application provides some freedom to define their activities and goals, motivators, and barriers, along with the structured alternatives. The structured parameters are embedded to find themes of concerns, activities targeted for behavior change, and for measuring outcomes and trajectories of change from a public health perspective. The purpose is also to generate supporting and challenging arguments based on momentary assessments, as well as analyzes of activities over time.

The exploration of participants' views on roles and behaviors of a digital agent in the context of supporting behavior change for improving health generated the framework for outlining an agent's emotional *support* and *challenge* in relation to the agent's role and the user's stage of change. We exemplify how the agent can take on behaviors and roles and shift between these by using

argumentation schemes. To encompass also the emotional support and challenge, two schemes for the purpose were defined to complement the schemes outlined by Walton et al. (2008). We built new schemes for the two and showed their usage through an example. In connection with the two new schemes, two new positions, *Position to Support* and *Position to Create Tension*, were introduced. While support and challenge is embedded in the argumentation frameworks' attack and support relations, there is currently no usage of such argumentation schemes through a multi-charactered digital companion for improving health, as far as we are aware of. This approach allows for managing arguments that have both emotion-based grounds and knowledge-based grounds, for instance, medical knowledge.

Our approach provides means to reason also about the ethical aspects in a dialogue situation which may trigger cognitive dissonance, which in turn, for some individuals, may increase anxiety and stress (Tengland, 2016). Guided by the domain experts' and participants' perspectives, the user's preferences are embedded in the two argumentation schemes as the representation of the mutual agreement on how the collaborative relationship should be actuated in terms of support and challenge creating tension. Furthermore, allowing the user to raise the topic of how to act as the third type of response paves also ways to allow the user to challenge the agent's behavior.

From a foundational argumentation perspective, it is worth highlighting that the results hint at the relevance of "soft" and informal behavioral and interactive properties of argumentation-augmented agents. In particular, our study results indicate that the preferred properties, e.g., regarding how *challenging* an agent is (which can, in our context, be interpreted as how consequent and with which attitude an agent will attempt to persuade with rational arguments), are subjective. Although these observations are not particularly surprising in their preliminary nature, it is worth noting that very little is known about human attitudes regarding the behavior of agents that have been augmented with (formal) argumentative capabilities. Even on object level, when assessing the inference results provided by abstract argumentation semantics, a recent study shows that the expectations of non-expert humans are not aligned with the behavior of many argumentation semantics that is popular in the research community (Guillaume et al., 2022). There seems to be little work that systematically studies how meta-level properties of computational argumentation, such as the way arguments or argumentation-based inferences are rendered to human users by a user interface, affect credibility, persuasiveness, and engagement. Considering the widespread success of choice architecture (Thaler et al., 2013) (also referred to as *nudging*), i.e., the rendering of information in a way that maximizes the intended impact on information consumers, this raises the question whether future approaches to argumentation for human-AI interaction can potentially benefit from fusing formal ("hard"), object-level argumentation with informal ("soft"), meta-level optimization, and personalization.

To summarize, our approach using computational argumentation and argument schemes provides transparency with respect to the agent's roles, behaviors, and sources of its arguments. Future user studies will explore how the user relates to the roles and positions of the agent in situated activities and the agent's support in the pursuit of improved health in these situations.

5.1. Participants' perceptions of emotional support vs. challenge

Since the results did not provide clear patterns of preferences among roles and behaviors relating to which TTM stage a user may be in, we choose to rely on the individual user's preferences, together with suggestions provided by the domain experts on how to address individuals in different stages of readiness for change.

An interesting observation was that the participants perceived neutral behavior as friendly and empathic in the situation when the human expressed distress due to overload at work. This occurred when the persona in the scenario shifted from the first one focusing on physical activity to the persona dealing with stress and worries. Their perception of the neutral agent as being empathic and friendly may be due to this kind of behavior is expected in such situations, and consequently, the participants interpret the agent's neutral behavior as such. One could also expect that the participants would have experienced a lack of empathy in this situation, as some participants expressed in a study on humans interacting with a robot (Tewari and Lindgren, 2022). However, as argued by Pulman (2010): "... a Companion which behaved in the same way whatever our emotional state would be thought of as insufficiently aware of us. But this may not mean that the Companion itself has to express emotions: all that is necessary to achieve this is the ability to recognize our own displays of emotion."

In the three cases when there was a difference between the intended character and behavior and how the participants rated the agent's behavior, the difference mainly consisted in that the participants rated the agent's empathy and friendliness higher than was intended, which also led to classifying these agents being companions to a larger extent. This we interpret as a cultural aspect; the participants were located in Scandinavia, where the way to express empathy and friendliness may differ from other places, a phenomenon which has been recently studied from an affective agents' perspective (Taverner et al., 2020). We plan to broaden our subsequent studies to include participants of various backgrounds to test our interpretation's validity.

People rate the high importance of changing behavior to decrease stress and tended to prefer a digital companion over other roles. This aligns with the expectation of a more empathic response in the exemplified dialogue on managing stress.

An outcome from the responses obtained from the participants for the question of which agent role they preferred in studies 1 and 2 was that more than 75% of them did not choose the companion role. On the other hand, the domain experts, although few, who experienced the dialogues with the digital agent through the prototype preferred the friendly and empathic role more than the other roles. The participants in studies 1 and 2 answered this question before they had encountered the scenarios and may have had a different view after evaluating the scenarios or if they had experienced the dialogues as the participating experts did through the prototype. Future studies will provide hands-on experiences of the different roles, which is expected to provide more reliable results.

The group of participants contained a large proportion of 30–39-year-old people in studies 1 and 2. It would be interesting to further analyze the data to explore whether the preferences that the group as a whole differ when studying the aspects from the perspective of age groups.

Studies on preferences regarding agent characters have shown that age is a deciding factor when it comes to choosing a digital companion. For instance, in [Hurmuz et al. \(2022\)](#), older adults preferred personalized content when interacting with a digital companion. Furthermore, when looking at the features of a digital companion in terms of friendliness, expertise, reliability, authority, and involvement, the general and elderly population preferred a gendered digital companion, specifically a young female ([ter Stal et al., 2019](#)). As for the type of messages, users would like to receive from such technologies, it has been found that reports about progress, sent at the right time, rather than something educational, is preferable ([Klaassen et al., 2013](#)). It is important to highlight, however, that there is currently a lack of studies on the preferences of roles and behaviors of digital companions in the domain of behavior change. Our ongoing and future study includes extending and implementing tailored dialogue capabilities of the digital companion. User studies will be conducted to further explore how participants in different stages of change and with different preferences relate to the digital agent in real-life settings. Furthermore, the effects of having argument-based dialogues with the digital companion on users' attitudes toward and actual changes of behavior, as well as wellbeing, will be studied in a randomized control trial over 6 months and continued use additional 6 months.

6. Conclusion

The studies presented in this article have explored the roles that digital companions can play in supporting behavior changes, and the attitudes that users, as well as domain experts from different disciplines, have toward them. A focus was placed on argumentative approaches, both conceptually, i.e., expectations and perceptions regarding the argumentation-related behavior and interaction, and practically, in the forms of argumentation-based system architecture and an early-stage prototype. The findings provide initial quantitative and qualitative insights that highlight the importance of “soft” non-formal behavioral aspects of argumentation-augmented agents in human-AI interaction scenarios but also indicate that some of the desirable properties of these aspects can be subjective and context-dependent.

Assuming that a major purpose of computational argumentation is the facilitation of human-machine interaction, we hence conclude that a nascent, high-potential research focus of the human-centered AI community in general, and the argumentation community in particular, could be the integration of “rational” argumentation-based reasoning by computational means with human-centered approaches regarding the presentation of arguments and argumentation-based inference results. To advance this research direction, results and methods from adjacent disciplines, such as behavioral economics and psychology, need to be incorporated. In turn, these disciplines can potentially—given that such an integration succeeds—benefit from the computational tools that the argumentation community provides.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Ethics statement

The project was reviewed and approved by the Swedish Ethical Review Authority (Dnr: 2019-02924 and Dnr: 2020-02985). The participants provided their written informed consent to participate in this study.

Author contributions

HL, KK, and SW: idea development and studies 1 and 2. KK: led the authoring, reviewed related work, and major work on the results relating to the person-tailored argument-based micro-dialogues. KK and HL: study 3 and development of the agent dialogue demonstrator. HL, SW, and TK: edit and review. HL: initial ideas, an overall responsibility of studies, the ACKTUS platform, and STAR-C application. All authors contributed to the article and approved the submitted version.

Funding

Research was partially funded by the Marianne and Marcus Wallenberg Foundation (Dnr MMW 2019.0220), and Wallenberg AI, Autonomous Systems and Software Program-Humanity and Society (WASP-HS). Further, the research programme grant from Forte, the Swedish Research Council for Health, Working Life and Welfare, supports STAR-C during 2019-2024 (Dnr. 2018-01461). This work was also partially funded by The Humane-AI-Net excellence network funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 952026.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2023.1069455/full#supplementary-material>

References

- Amgoud, L., and Cayrol, C. (2013). On the acceptability of arguments in preference-based argumentation. *arXiv preprint arXiv:1301.7358*. doi: 10.48550/arXiv.1301.7358
- Amgoud, L., Cayrol, C., Lagasque-Schiek, M.-C., and Livet, P. (2008). On bipolarity in argumentation frameworks. *Int. J. Intell. Syst.* 23, 1062–1093. doi: 10.1002/int.20307
- Atkinson, K., Bench-Capon, T., and McBurney, P. (2006). Computational representation of practical argument. *Synthese* 152, 157–206. doi: 10.1007/s11229-005-3488-2
- Atkinson, K., and Wyner, A. (2013). “The value of values in computational argumentation,” in *From Knowledge Representation TO Argumentation in AI, Law and Policy Making: A Festschrift in Honour of Trevor Bench-Capon on the Occasion of his 60th Birthday* (University of Liverpool), 39–62.
- Ballnat, S., and Gordon, T. F. (2010). “Goal selection in argumentation processes,” in *Computational Models of Argument: Proceedings of COMMA Vol. 2010* (IOS Press), 51. doi: 10.3233/978-1-60750-619-5-51
- Baskar, J., Janols, R., Guerrero, E., Nieves, J. C., and Lindgren, H. (2017). “A multipurpose goal model for personalised digital coaching,” in *Agents and Multi-Agent Systems for Health Care* (Cham: Springer International Publishing), 94–116. doi: 10.1007/978-3-319-70887-4_6
- Bench-Capon, T. (2002). Value based argumentation frameworks. *arXiv preprint cs/0207059*. doi: 10.48550/arXiv.cs/0207059
- Bench-Capon, T. J., Doutre, S., and Dunne, P. E. (2007). Audiences in argumentation frameworks. *Artif. Intell.* 171, 42–71. doi: 10.1016/j.artint.2006.10.013
- Blakey, J., and Day, I. (2012). *Challenging Coaching: Going Beyond Traditional Coaching to Face the FACTS*. Boston, MA: Nicolas Brealey Publishing.
- Blomstedt, Y., Norberg, M., Stenlund, H., Nyström, L., Lönnberg, G., Boman, K., et al. (2015). Impact of a combined community and primary care prevention strategy on all-cause and cardiovascular mortality: a cohort analysis based on 1 million person-years of follow-up in västerbotten county, Sweden, during 1990–2006. *BMJ Open* 5, e009651. doi: 10.1136/bmjopen-2015-009651
- Braun, F., Block, L., and Stegmüller, S. (2021). “Josy: development of a digital companion for elderly people—a new way to experience technology,” in *International Conference on Applied Human Factors and Ergonomics* (Springer), 436–442.
- Chalaguine, L. A., Hunter, A., Potts, H., and Hamilton, F. (2019). “Impact of argument type and concerns in argumentation with a chatbot,” in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)* (Portland, OR: IEEE), 1557–1562.
- Chesñevar, C., McGinnis, J., Modgil, S., Rahwan, I., Reed, C., Simari, G., et al. (2006). Towards an argument interchange format. *Knowl. Eng. Rev.* 21, 293–316. doi: 10.1017/S0269888906001044
- Çyras, K., Rago, A., Albini, E., Baroni, P., and Toni, F. (2021). “Argumentative XAI: a survey,” in *30th International Joint Conference on Artificial Intelligence*, ed Z.-H. Zhou (Montreal: IJCAI), 4392–4399.
- De Boni, M., Hurling, R., and Dryden, W. (2006). “Argumentation through an automated rational-emotive behavior therapy system for change in exercise behavior,” in *AAAI Spring Symposium: Argumentation for Consumers of Healthcare*, 34–38.
- Dietz, E., and Kakas, A. (2021). “Cognitive argumentation and the selection task,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 43, 1588–1594.
- Donadello, I., Hunter, A., Teso, S., and Dragoni, M. (2022). Machine learning for utility prediction in argument-based computational persuasion. *Proc. AAAI Conf. Artif. Intell.* 36, 5592–5599. doi: 10.1609/aaai.v36i5.20499
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77, 321–357. doi: 10.1016/0004-3702(94)00041-X
- Gilbert, P. (2009). Introducing compassion-focused therapy. *Adv. Psychiatr. Treat.* 15, 199–208. doi: 10.1192/apt.bp.107.005264
- Grasso, F., Cawsey, A., and Jones, R. B. (2000). Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition. *Int. J. Hum. Comput. Stud.* 53, 1077–1115. doi: 10.1006/ijhc.2000.0429
- Guerrero, E., and Lindgren, H. (2021a). “Causal interventions with formal argumentation theory,” in *LAMAS and LR, 2021. Proceedings LAMAS and LR 2021*. doi: 10.1007/978-3-030-85710-3_13
- Guerrero, E., and Lindgren, H. (2021b). “Towards motivation-driven intelligent interfaces: formal argumentation meets activity theory,” in *2021 Swedish Artificial Intelligence Society Workshop (SAIS)*, 1–4. doi: 10.1109/SAIS53221.2021.9484008
- Guillaume, M., Cramer, M., van der Torre, L., and Schiltz, C. (2022). Reasoning on conflicting information: an empirical study of formal argumentation. *PLoS ONE* 17, e0273225. doi: 10.1371/journal.pone.0273225
- Hadoux, E., and Hunter, A. (2019). Comfort or safety? gathering and using the concerns of a participant for better persuasion. *Argument Computa.* 10, 113–147. doi: 10.3233/AAC-191007
- Hadoux, E., Hunter, A., and Polberg, S. (2018). “Biparty decision theory for dialogical argumentation,” in *Computational Models of Argument* (IOS Press), 233–240. doi: 10.3233/978-1-61499-906-5-233
- Hörnsten, Å. A., Lindahl, K. B., Persson, K. I., and Edvardsson, K. (2014). Strategies in health-promoting dialogues—primary healthcare nurses’ perspectives—a qualitative study. *Scand. J. Caring Sci.* 28, 235–244. doi: 10.1111/scs.12045
- Hurmuz, M., Jansen Kosterink, S., Beinema, T., Fischer, K., Akker, H., and Hermens, H. (2022). Evaluation of a virtual coaching system health intervention: a mixed methods observational cohort study in the Netherlands. *Internet Intervent.* 27, 100501. doi: 10.1016/j.invent.2022.100501
- Jang, J., and Kim, J. (2020). Healthier life with digital companions: effects of reflection-level and statement-type of messages on behavior change via a perceived companion. *Int. J. Hum. Comput. Interact.* 36, 172–189. doi: 10.1080/10447318.2019.1615722
- Kantharaju, R. B., Pease, A., Reidsma, D., Pelachaud, C., Snaith, M., Bruijnes, M., et al. (2019). “Integrating argumentation with social conversation between multiple virtual coaches,” in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 203–205. doi: 10.1145/3308532.3329450
- Kaptelinin, V., and Nardi, B. A. (2006). *Acting With Technology: Activity Theory and Interaction Design*. MIT Press.
- Klaassen, R., Akker, R., Lavrysen, T., and Wissen, S. (2013). User preferences for multi-device context-aware feedback in a digital coaching system. *J. Multimodal User Interfaces* 7, 247–267. doi: 10.1007/s12193-013-0125-0
- Kobsa, A. (1990). User modeling in dialog systems: potentials and hazards. *AI Soc.* 4, 214–231. doi: 10.1007/BF01889941
- Latham, G. P., and Locke, E. A. (1991). Self-regulation through goal setting. *Organ. Behav. Hum. Decis. Process.* 50, 212–247. doi: 10.1016/0749-5978(91)90021-K
- Lindgren, H., Guerrero, E., Jingar, M., Lindvall, K., Ng, N., Richter Sundberg, L., et al. (2020). The STAR-C intelligent coach: a cross-disciplinary design process of a behavior change intervention in primary care. *Stud. Health Technol. Inform.* 273, 203–208. doi: 10.3233/SHTI200640
- Lindgren, H., and Weck, S. (2021). Conceptual model for behaviour change progress - instrument in design processes for behaviour change systems. *Stud. Health Technol. Inform.* 285, 277–280. doi: 10.3233/SHTI210614
- Lindgren, H., and Weck, S. (2022). “Contextualising goal setting for behavior change—from baby-steps to value directions,” in *ACM European Conference on Cognitive Ergonomics (ECCE)*, Vol. 2022.
- Lindgren, H., and Yan, C. (2015). “ACKTUS: a platform for developing personalized support systems in the health domain,” in *Proceedings of the 5th International Conference on Digital Health 2015, DH '15* (New York, NY: ACM), 135–142.
- Lindholm, L., Stenling, A., Norberg, M., Stenlund, H., and Weinehall, L. (2018). A cost-effectiveness analysis of a community based cvd program in sweden based on a retrospective register cohort. *BMC Public Health* 18, 452. doi: 10.1186/s12889-018-5339-3
- Locke, E. A., and Latham, G. P. (1984). *Goal Setting: A Motivational Technique That Works!* Prentice Hall.
- Lunenburg, F. C. (2011). Goal-setting theory of motivation. *Int. J. Manag. Bus. Administ.* 15, 1–6.
- Manning, J. B., Blandford, A., and Edbrooke-Childs, J. (2022). Digital companion choice to support teachers’ stress self-management: systematic approach through taxonomy creation. *JMIR Format. Res.* 6, e32312. doi: 10.2196/32312
- Miller, W. R., and Rollnick, S. (2012). *Motivational Interviewing: Helping People Change*. Guilford Press.
- Ng, N., Eriksson, M., Guerrero, E., Gustafsson, C., Kinsman, J., Lindberg, J., et al. (2021). Sustainable behavior change for health supported by person-tailored, adaptive, risk-aware digital coaching in a social context: study protocol for the star-c research program. *Front. Public Health* 9, 138. doi: 10.3389/fpubh.2021.593453
- Nguyen, H., and Masthoff, J. (2008). “Designing persuasive dialogue systems: using argumentation with care,” in *International Conference on Persuasive Technology* (Springer), 201–212. doi: 10.1007/978-3-540-68504-3_18
- Nowak, A., Lukowicz, P., and Horodecki, P. (2018). Assessing artificial intelligence for humanity: will ai be the our biggest ever advance? or the biggest threat [opinion]. *IEEE Technol. Soc. Mag.* 37, 26–34. doi: 10.1109/MTS.2018.2876105
- op den Akker, H., Jones, V., and Hermens, H. J. (2014). Tailoring real-time physical activity coaching systems: a literature survey and model. *User Model Useradapt Interact.* 24, 351–392. doi: 10.1007/s11257-014-9146-y
- Perelman, C., and Olbrechts-Tyteca, L. (1969). *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press.
- Prochaska, J., Redding, C., and Evers, K. (2015). “The transtheoretical model and stages of change,” in *Health Behavior: Theory, Research, and Practice*, eds K. Glanz, B. K. Rimer, and K. V. Viswanath (Jossey-Bass/Wiley), 60–84.
- Pulman, S. (2010). “Conditions for companionship,” in *Close Engagements With Artificial Companions-Key Social, Psychological, Ethical and Design Issues* (Philadelphia, PA: John Benjamins Publishing Company), 29–34.
- Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020

- Spirig, J., Garcia, K., and Mayer, S. (2021). "An expert digital companion for working environments," in *11th International Conference on the Internet of Things*, 25–32. doi: 10.1145/3494322.3494326
- Steels, L. (2020). "Personal dynamic memories are necessary to deal with meaning and understanding in human-centric AI," in *NeHuAI@ECAI* (CEUR-WS.org), 11–16.
- Taverner, J., Vivancos, E., and Botti, V. (2020). A multidimensional culturally adapted representation of emotions for affective computational simulation and recognition. *IEEE Tran. Affect. Comput.* 1–10. doi: 10.1109/TAFFC.2020.3030586
- Tengland, P. (2016). Behavior change or empowerment: on the ethics of health-promotion goals. *Health Care Anal.* 24, 24–46. doi: 10.1007/s10728-013-0265-0
- ter Stal, S., Tabak, M., Akker, H., Beinema, T., and Hermens, H. (2019). Who do you prefer? the effect of age, gender and role on users' first impressions of embodied conversational agents in ehealth. *Int. J. Hum. Comput. Interact.* 36, 1–12. doi: 10.1080/10447318.2019.1699744
- Tewari, M., and Lindgren, H. (2022). Expecting, understanding, relating and interacting - older and younger adults' perspectives on breakdown situations in human-robot dialogues. *Front. AI Robot.* 9, 956709. doi: 10.3389/frobt.2022.956709
- Thaler, R. H., Sunstein, C. R., and Balz, J. P. (2013). *Choice Architecture*, Vol. 2013. Princeton, NJ: Princeton University Press.
- Torous, J., Nicholas, J., Larsen, M. E., Firth, J., and Christensen, H. (2018). Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *Evid. Based Ment. Health* 21, 116–119. doi: 10.1136/eb-2018-102891
- van der Weide, T. L. (2011). *Arguing to motivate decisions* (Ph.D. thesis). Utrecht University.
- Vassiliades, A., Bassiliades, N., and Patkos, T. (2021). Argumentation and explainable artificial intelligence: a survey. *Knowl. Eng. Rev.* 36, e5. doi: 10.1017/S0269888921000011
- Walton, D. (1996). *Argumentation Schemes for Presumptive Reasoning*. London: Routledge.
- Walton, D., and Krabbe, E. C. W. (1995). *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. New York, NY: State University of New York Press.
- Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge: Cambridge University Press.
- Weber, F., Schrumpf, J., and Thelen, T. (2021). "Development of a digital goal setting companion for higher education," in *DELFI 2021* (Gesellschaft für Informatik eV).
- Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R. J., et al. (2018). The grand challenges of science robotics. *Sci. Robot.* 3, eaar7650. doi: 10.1126/scirobotics.aar7650



OPEN ACCESS

EDITED BY

Antonis Kakas,
University of Cyprus, Cyprus

REVIEWED BY

Markus Ulbricht,
Leipzig University, Germany
Johannes Wallner,
Graz University of Technology, Austria

*CORRESPONDENCE

Marcos Cramer
✉ marcos.cramer@tu-dresden.de

SPECIALTY SECTION

This article was submitted to
Machine Learning and Artificial Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

RECEIVED 15 September 2022

ACCEPTED 10 February 2023

PUBLISHED 23 March 2023

CITATION

Cramer M and van der Torre L (2023) An
argumentation semantics for rational human
evaluation of arguments.
Front. Artif. Intell. 6:1045663.
doi: 10.3389/frai.2023.1045663

COPYRIGHT

© 2023 Cramer and van der Torre. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

An argumentation semantics for rational human evaluation of arguments

Marcos Cramer^{1*} and Leendert van der Torre²

¹Institute for Artificial Intelligence, TU Dresden, Dresden, Germany, ²Department of Computer Science, University of Luxembourg, Esch-sur-Alzette, Luxembourg

In abstract argumentation theory, many argumentation semantics have been proposed for evaluating argumentation frameworks. This article is based on the following research question: Which semantics corresponds well to what humans consider a rational judgment on the acceptability of arguments? There are two systematic ways to approach this research question: A normative perspective is provided by the principle-based approach, in which semantics are evaluated based on their satisfaction of various normatively desirable principles. A descriptive perspective is provided by the empirical approach, in which cognitive studies are conducted to determine which semantics best predicts human judgments about arguments. In this article, we combine both approaches to motivate a new argumentation semantics called SCF2. For this purpose, we introduce and motivate two new principles and show that no semantics from the literature satisfies both of them. We define SCF2 and prove that it satisfies both new principles. Furthermore, we discuss findings of a recent empirical cognitive study that provide additional support to SCF2.

KEYWORDS

knowledge representation, formal argumentation, abstract argumentation, argumentation semantics, principle-based approach

1. Introduction

The formal study of argumentation is an important field of research within AI (Rahwan and Simari, 2009), in particular in the area of knowledge representation and reasoning, and in the area of multiagent systems. Argumentation as inference provides a general framework for non-monotonic reasoning, and argumentation as dialogue provides a general framework for agent interaction (Prakken, 2018). Argumentation-based approaches are assumed to be better suited for modeling human reasoning than traditional logical methods used in knowledge representation and reasoning, including reasoning in the context of conflicting information and dealing with fallacies and other errors in human reasoning. Formal argumentation is a kind of argument reasoning and is often contrasted with other recent developments in computational argumentation in AI (Van Eemeren and Verheij, 2018), such as approaches based on argument mining (Budzynska and Villata, 2018; Lawrence and Reed, 2020), argument assessment, argument generation, and cognitive modeling (Lauscher et al., 2021).

A central focus of the modern development of formal argumentation has been the idea of Dung (1995) that under some conditions, the acceptance of arguments depends only on a so-called *attack* relation among the arguments, and not on the internal structure of the arguments. Dung called this approach *abstract* argumentation and called the directed graph that represents the arguments and the attack relation between them an *argumentation framework* (AF). Whether an argument is deemed acceptable depends on the decision about

other arguments. Therefore, the basic concept in abstract argumentation is a *set* of arguments that can be accepted together, called an *extension*. Crucially, there may be several of such extensions, and these extensions may be incompatible. An *extension-based argumentation semantics* takes as input an AF and produces as output a set of extensions.

Traditionally, two classes of extension-based argumentation semantics have been studied (Baroni et al., 2018). Dung introduced several examples of so-called *admissibility-based* semantics, formalizing the idea that an argument is acceptable in the context of an extension if the extension *defends* the argument, i.e., attacks all the attackers of the argument. In this article, we consider his grounded, complete, preferred, and stable semantics. Moreover, we consider the admissibility-based semantics known as semi-stable semantics (Verheij, 1996; Caminada et al., 2012). The other kind of extension-based argumentation semantics is *naive-based* semantics, which is based on the idea that acceptable argument sets are specific maximal conflict-free sets. In this article, we consider the naive, stage, CF2 and stage2 semantics and develop a new naive-based semantics called SCF2. More recently, some semantics have been introduced that are neither admissibility based nor naive based (Dvorák et al., 2022); see the related work section of this article for further details.

Abstract argumentation has various potential applications (Rahwan and Simari, 2009), and the choice of semantics depends on the envisioned application. In this article, we focus on the following research question: Which semantics corresponds well to what humans consider a rational judgment on the acceptability of arguments?

There are two systematic ways to approach this research question: A normative perspective is provided by the *principle-based approach* (Baroni and Giacomin, 2007; van der Torre and Vesic, 2018), in which semantics are evaluated based on their satisfaction of various normatively desirable principles. A descriptive perspective is provided by the *empirical approach* (Rahwan et al., 2010), in which cognitive studies are conducted to determine which semantics best predicts human judgments about arguments. In this article, we combine both approaches.

Two recent empirical cognitive studies on argumentation semantics (Cramer and Guillaume, 2018b, 2019) showed CF2 to be better predictors of human argument evaluation than admissibility-based semantics like grounded and preferred. This finding sheds some doubt on principles that are only satisfied by admissibility-based semantics, e.g., admissibility, defense, and reinstatement, as surveyed by van der Torre and Vesic (2018). For this reason, in this article, we focus on other existing principles (e.g., directionality) and introduce new ones.

The first new principle we consider is *irrelevance of Necessarily Rejected Arguments* (INRA). Informally, INRA says that if an argument is attacked by every extension of an AF, then deleting this argument should not change the set of extensions. The idea, here, is that an argument that is attacked by every extension would be rejected by any party in a debate and hence would never be brought up in a debate. Hence, it should be treated as if it did not even exist.

The second principle that we consider is *Strong Completeness Outside Odd Cycles* (SCOOCs). Informally, SCOOC says that if an argument a and its attackers are not in an odd cycle, then an

extension not containing any of a 's attackers must contain a . The principle is based on the idea that it is generally desirable that an argument that is not attacked by any argument in a given extension should itself be in that extension. While it is possible to ensure this property in AFs without odd cycles, this is not the case for AFs involving an odd cycle. The idea behind the SCOOC principle is to still satisfy this property as much as possible, i.e., whenever the argument under consideration and its attackers are not in an odd cycle.

We show that of the nine common semantics mentioned earlier, the only ones that satisfy INRA are grounded, complete and naive semantics. In addition, we show that a variant of CF2 that we call *nsa*(CF2) and that consists of first deleting all self-attacking arguments and then applying CF2 semantics also satisfies INRA.

Furthermore, we show that of these 10 semantics (the nine mentioned at the beginning and *nsa*(CF2)), the only one that satisfies SCOOC is the stable semantics. However, stable semantics satisfies neither directionality nor INRA. The fact that none of the considered existing semantics satisfies both new principles introduced in this article raises the question whether these two principles can be satisfied in conjunction. We answer this question positively by defining a novel semantics called *SCF2 semantics* that satisfies both of them.

Finally, we discuss the findings of a recent cognitive study by Cramer and Guillaume (2019) and observe that SCF2 explains the judgments of participants in this study better than any existing semantics. This provides additional support for our claim that SCF2 corresponds well to what humans consider a rational judgment on the acceptability of arguments.

This article is an extended version of a workshop article (Cramer and van der Torre, 2019). Compared to the workshop article, here, we give more background on the relation to human-centric AI and consider much more principles from the abstract argumentation literature: While in the workshop article, we focused on three principles, this article evaluates the new semantics against 37 principles. Furthermore, unlike in the workshop article, we give full proofs for all theorems that we present.

1.1. Relation to human-centric intelligence

Humans use arguments both as a means to persuade others in a dialogue and as a way to make decisions and draw tentative conclusions by comparing arguments for and against various positions. In order for AI technology to interact meaningfully with humans, argumentation as practiced by humans, therefore, needs to be taken into account.

Argumentation and dialogue have been studied in many fields. In artificial intelligence, a distinction can be made between formal argumentation and computational argumentation, where formal argumentation is concerned both with argumentation as inference studied in knowledge representation and reasoning and argumentation as dialogue studied in multiagent systems (Prakken, 2018). Since the work of Dung (1995), these approaches are studied not only at a logical or structured level but also at an abstract level.

Formal argumentation can be seen as a natural successor of logic-based approaches studied in the previous century (Prakken, 2018; Van Eemeren and Verheij, 2018). Approaches to human reasoning based on classical logic have little to say in case of conflict. However, humans need to reason about conflict all the time, for example, when receiving contradictory or false information or when dealing with opposing opinions. Formal argumentation goes beyond classical logic by presenting distinct rational viewpoints in case of conflict and by incorporating methods from non-monotonic logic to resolve some of these conflicts. They do this by modeling facts as assumptions and modeling rules as defeasible inferences. On the other hand, formal argumentation builds on traditional logical methods by representing the structure of individual arguments themselves in a logical way. Each extension of a set of acceptable arguments may be seen as a coherent viewpoint.

In the current article, the focus is on argumentation as inference and on abstract argumentation, the study of the relation among arguments with a focus on how the attack relation between arguments (when one argument is a counterargument to another) can serve as a basis for judgments about the acceptability of arguments. It can be seen as the study of a dialogue state at a single moment in time. Even when an argument is not accepted in any extension and thus can be ignored according to the INRA principle, the same argument can play a role later in the dialogue when the framework has changed.

Dung's theory is based on the assumption that the acceptance of arguments depends only on the attack relation among the constructive arguments, not on their internal structure. Dung's theory can be defended in different ways. Suppose the assumption is false, i.e., one of the dialogue participants believes that due to the internal structure of argument *A*, it cannot be accepted. Now suppose that another dialogue participant disagrees with this position and claims that the internal structure of the argument is completely fine. In this disagreement, we can model this disagreement with arguments *B* and *C* and the relation between arguments *A* and *B* with an attack from *B* to *A*. In general, the fact that in abstract argumentation, everything has to be modeled by an argument can be interpreted as the statement that every criticism can be criticized itself as well.

The methods of abstract argumentation are also relevant for the study of the internal structure of arguments and the dynamics of dialogue scenarios. When the internal structure of arguments is made explicit, and the arguments are attributed to the agents that put them forward, one can address how arguments are generated in light of other arguments and how that can lead to a resolution of conflicts and paradoxes. In such cases, the argumentation framework can change over time due to agent interaction.

Human reasoning is inherently non-monotonic: It often happens that one draws a conclusion from certain given information but later gives up that conclusion due to novel information speaking against it. This non-monotonicity of human reasoning cannot be modeled in classical monotonic logic. For this reason, non-monotonic logic has been designed since the 1980s. Since its inception in the early 1990s, formal argumentation has had a strong connection to non-monotonic logic. The idea, here, is that novel information allows us to construct new arguments, some of

which may attack previously accepted arguments and lead to their rejection. Thus, formal argumentation can often be viewed as a tool for making the inference process of non-monotonic logics explicit, concrete, and close in nature to actual human reasoning.

While some of the research in formal argumentation is somewhat detached from the human practice of argumentation, there are also many researchers who aim at building a bridge between human reasoning and formal argumentation by studying how various formalisms and semantics from formal argumentation relate to actual human reasoning. For example, formal argumentation has been combined with approaches based on natural language processing and argument mining (Budzynska and Villata, 2018). Furthermore, as detailed in Section 6, multiple cognitive studies have been conducted to investigate the relation between human reasoning and argumentation formalisms.

With the help of such interdisciplinary research, formal argumentation is becoming more relevant to the endeavor of human-centric AI. This article aims to contribute to this research by studying which argumentation semantics (i.e., which method for evaluating the acceptability of arguments based on the attack relation between the arguments) is a good model for rational human evaluation of arguments. For this, two approaches are combined as follows:

- A normative perspective is provided by the principle-based approach, in which semantics are evaluated based on their satisfaction of various normatively desirable principles.
- A descriptive perspective is provided by the empirical approach, in which cognitive studies are conducted to determine which semantics best predicts human judgments about arguments.

In this article, we argue that the SCF2 semantics is a reasonable choice from both points of view. It may thus be better suited for human-centric AI than other argumentation semantics proposed in the literature.

2. Preliminaries

In this section, we define required notions from abstract argumentation theory Dung (1995) and Baroni et al. (2018). In addition, we define three principles from the literature on principle-based argumentation (Baroni and Giacomin, 2007; van der Torre and Vesic, 2018) and present an argument for the case that the directionality principle is a desirable property for a semantics designed to match what humans would consider a rational judgment on the acceptability of arguments.

DEFINITION 1. An *argumentation framework* (AF) $F = \langle Ar, att \rangle$ is a finite directed graph in which the set *Ar* of vertices is considered to represent arguments and the set *att* of edges is considered to represent the attack relation between arguments, i.e., the relation between a counterargument and the argument that it counters.

DEFINITION 2. An *att-path* is a sequence $\langle a_0, \dots, a_n \rangle$ of arguments where $(a_i, a_{i+1}) \in att$ for $0 \leq i < n$ and where $a_j \neq a_k$

for $0 \leq j < k \leq n$ with either $j \neq 0$ or $k \neq n$. An *odd att-cycle* is an *att-path* $\langle a_0, \dots, a_n \rangle$ where $a_0 = a_n$ and n is odd.

DEFINITION 3. Let $F = \langle Ar, att \rangle$ be an AF, and let $S \subseteq Ar$. We write $F|_S$ for the restricted AF $\langle S, att \cap (S \times S) \rangle$. The set S is called *conflict-free* iff there are no arguments $b, c \in S$ such that b attacks c (i.e., such that $(b, c) \in att$). Argument $a \in Ar$ is *defended* by S iff for every $b \in Ar$ such that b attacks a there exists $c \in S$ such that c attacks b . We say that S *attacks* a if there exists $b \in S$ such that b attacks a , and we define $S^+ = \{a \in Ar \mid S \text{ attacks } a\}$ and $S^- = \{a \in Ar \mid a \text{ attacks some } b \in S\}$.

- S is a *complete extension* of F iff it is conflict free, it defends all its arguments, and it contains all the arguments it defends.
- S is a *stable extension* of F iff it is conflict free, and it attacks all the arguments of $A \setminus S$.
- S is the *grounded extension* of F iff it is a minimal with respect to set inclusion complete extension of F .
- S is a *preferred extension* of F iff it is a maximal with respect to set inclusion complete extension of F .
- S is a *semi-stable extension* of F iff it is a complete extension, and there exists no complete extension S_1 such that $S \cup S^+ \subset S_1 \cup S_1^+$.
- S is a *stage extension* of F iff S is a conflict-free set, and there exists no conflict-free set S_1 such that $S \cup S^+ \subset S_1 \cup S_1^+$.
- S is a *naive extension* of F iff S is a maximal conflict-free set.

CF2 semantics was first introduced by Baroni et al. (2005). The idea behind it is that we partition the AF into *strongly connected components* and recursively evaluate it component by component by choosing maximal conflict-free sets in each component and removing arguments attacked by chosen arguments. We formally define it following the notation of Dvořák and Gaggl (2016). For this, we first need some auxiliary notions:

DEFINITION 4. Let $F = \langle Ar, att \rangle$ be an AF, and let $a, b \in Ar$. We define $a \sim b$ iff either $a = b$ or there is an *att-path* from a to b , and there is an *att-path* from b to a . The equivalence classes under the equivalence relation \sim are called *strongly connected components* (SCCs) of F . We denote the set of SCCs of F by $SCCs(F)$. Given $S \subseteq Ar$, we define $D_F(S) := \{b \in Ar \mid \exists a \in S : (a, b) \in att \wedge a \not\sim b\}$.

If $F = \langle \emptyset, \emptyset \rangle$, we consider \emptyset to be an SCC of F ; else \emptyset is not an SCC.

The simplified SCC-recursive scheme used for defining CF2 and stage2 is a function that maps a semantics σ to another semantics $scc(\sigma)$:

DEFINITION 5. Let σ be an argumentation semantics. The argumentation semantics $scc(\sigma)$ is defined as follows. Let $F = \langle Ar, att \rangle$ be an AF, and let $S \subseteq Ar$. Then S is an $scc(\sigma)$ -extension of F iff either

- $|SCCs(F)| \leq 1$ and S is a σ -extension of F , or
- $|SCCs(F)| > 1$ and for each $C \in SCCs(F)$, $S \cap C$ is an $scc(\sigma)$ -extension of $F|_{C \setminus D_F(S)}$.

CF2 semantics is defined to be $scc(naive)$, and stage2 semantics is defined to be $scc(stage)$.

Apart from the function scc , we introduce a further function—called nsa —that also maps a semantics to another semantics. Informally, the idea behind $nsa(\sigma)$ is that we first delete all self-attacking arguments and then apply σ . To define nsa formally, we first need an auxiliary definition:

DEFINITION 6. Let $F = \langle Ar, att \rangle$ be an AF. We define the *non-self-attacking restriction* of F , denoted by $NSA(F)$, to be the AF $F_{Ar'}$, where $Ar' := \{a \in Ar \mid (a, a) \notin att\}$.

DEFINITION 7. Let σ be an argumentation semantics. The argumentation semantics $nsa(\sigma)$ is defined as follows. Let $F = \langle Ar, att \rangle$ be an AF, and let $S \subseteq Ar$. We say that E is an $nsa(\sigma)$ -extension of F iff E is a σ -extension of $NSA(F)$.

We now define the directionality principle introduced by Baroni and Giacomin (2007). For this, we first need an auxiliary notion:

DEFINITION 8. Let $F = \langle Ar, att \rangle$ be an AF. A set $U \subseteq Ar$ is *unattacked* iff there exists no $a \in A \setminus U$ such that a attacks some $b \in U$.

DEFINITION 9. A semantics σ satisfies the *directionality* principle iff for every AF F and every unattacked set U ; it holds that $\sigma(F|_U) = \{E \cap U \mid E \in \sigma(F)\}$.

The directionality principle corresponds to an important feature of the human practice of argumentation, namely that if a person has formed an opinion on some arguments and is confronted with new arguments, they will only feel compelled to reconsider their judgment on the prior arguments if one of the new arguments attacks one of the prior arguments. Apart from our own intuition, we can also refer to the results of an empirical cognitive study on argumentation that shows that humans are able to systematically judge the directionality of attacks between arguments (Cramer and Guillaume, 2018a). Thus, we consider the directionality principle crucial for the goal that we focus on in this article.

We define two further principles from the literature on principle-based argumentation (Baroni and Giacomin, 2007; van der Torre and Vesic, 2018) that are relevant for getting a better picture of the behavior of a semantics and can be used to derive multiple further principles proposed in the literature.

DEFINITION 10. A semantics σ satisfies the *naivety* principle if and only if for every AF F , for every $E \in \sigma(F)$, E is a maximal with respect to set inclusion conflict-free set in F .

DEFINITION 11. Given an argumentation framework $F = \langle Ar, att \rangle$ and sets $S, E \subseteq Ar$, we define $U_F(S, E) := \{a \in S \mid \nexists b : (b, a) \in att, b \not\sim a, \text{ and } E \text{ does not attack } b\}$.

DEFINITION 12. A binary function BF is called a *base function* iff for every AF $F = \langle Ar, att \rangle$ such that $|SCCs(F)| = 1$ and every set $C \subseteq Ar$, $BF(F, C) \subseteq \mathcal{P}(Ar)$.

Here the notation $\mathcal{P}(Ar)$ denotes the powerset of Ar , i.e., the set of all subsets of Ar .

DEFINITION 13. Given a base function BF , an AF $F = (Ar, att)$ and a set $C \subseteq Ar$, we recursively define $GF(BF, F, C) \subseteq \mathcal{P}(Ar)$ as follows: for every $E \subseteq Ar$, $E \in GF(BF, F, C)$ iff

- in case $|SCCs(F)| = 1$, $E \in BF(F, C)$,
- otherwise, for all $S \in SCCs(F)$, $(E \cap S) \in GF(BF, F|_{S \setminus D_F(E)}, U_F(S, E) \cap C)$.

DEFINITION 14. A semantics σ satisfies the *SCC-recursiveness* principle iff there is a base function BF such that for every AF $F = (Ar, att)$ we have $\sigma(F) = GF(BF, F, Ar)$.

3. Two new principles

In this section, we define and motivate the two new principles introduced in the article. Let us first look at the principle that we call *Irrelevance of Necessarily Rejected Arguments (INRA)*. The idea behind this principle is that in order for an argument to be relevant in a debate, there must be a coherent standpoint according to which this argument is accepted or at least not clearly rejected. If an argument is attacked by an extension, it would be clearly rejected by any rational person whose standpoint is described by the extension in question. So, if an argument is attacked by every extension, it is clearly rejected in light of every rational standpoint and would, therefore, never be brought up in a debate between rational people. For the purpose of evaluating the acceptability of arguments, it, therefore, makes sense to treat such an argument as if it did not even exist. Talking in the language of extensions, this can be formulated as follows: If an argument a is attacked by every extension of an AF, then deleting a should not change the set of extensions.¹

In order to formally define the INRA principle, we first need to define a notation for an AF with one argument deleted:

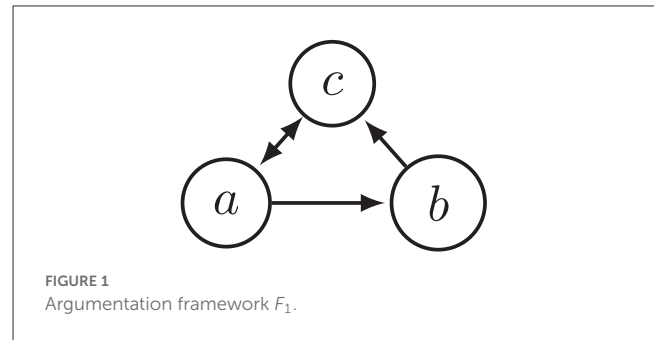
DEFINITION 15. Let $F = (Ar, att)$ be an AF and let $a \in Ar$ be an argument. Then F^{-a} denotes the restricted AF $F|_{Ar \setminus \{a\}}$.

DEFINITION 16. Let σ be an argumentation semantics. We say that σ satisfies *Irrelevance of Necessarily Rejected Arguments (INRA)* iff for every AF $F = (Ar, att)$ and every argument $a \in Ar$, if every $E \in \sigma(F)$ attacks a , then $\sigma(F) = \sigma(F^{-a})$.

We now illustrate the definition through an example of the preferred semantics:

EXAMPLE 1. Consider the argumentation framework F_1 depicted in Figure 1. The only preferred extension of F_1 is $\{a\}$. This extension attacks b . So, b is attacked by every extension of F_1 . If we remove argument b from F_0 , we are left with the AF F_1^{-b} consisting only of a and c attacking each other. F_1^{-b} has two preferred extensions,

¹ Note that the deletion of arguments mentioned in this principle only concerns the procedure for deciding which arguments are accepted according to the SCF2 argumentation semantics. In applications of the SCF2 semantics to structured argumentation or to the formal study of dialogues, the deletion of arguments would not happen at the level of argument construction but only at the level of argument evaluation. So, even arguments that are rejected by everyone could influence the dynamics of argument construction by participants of a dialogue.



$\{a\}$ and $\{c\}$. So, when removing an argument (namely b) that was attacked by every extension, the set of extensions changed. Thus, this example constitutes a violation of the INRA principle. We have, therefore, established that the preferred semantics does not satisfy INRA.

The second principle that we consider is *Strong Completeness Outside Odd Cycles (SCOOC)*. Informally, SCOOC says that if an argument a and its attackers are not in an odd cycle, then an extension not containing any of a 's attackers must contain a .

In order to formally define the Strong Completeness Outside Odd Cycles principle, we first need to define a notation for the set of all attackers of an argument and the auxiliary notion of a set of arguments being *strongly complete outside odd cycles*.

DEFINITION 17. Let $F = (Ar, att)$ be an AF, and let $A \subseteq Ar$. We say that A is *strongly complete outside odd cycles* iff for every argument $a \in Ar$, the following condition holds: If

- no argument in $\{a\} \cup \{a\}^-$ is in an odd *att-cycle*, and
- $A \cap \{a\}^- = \emptyset$,

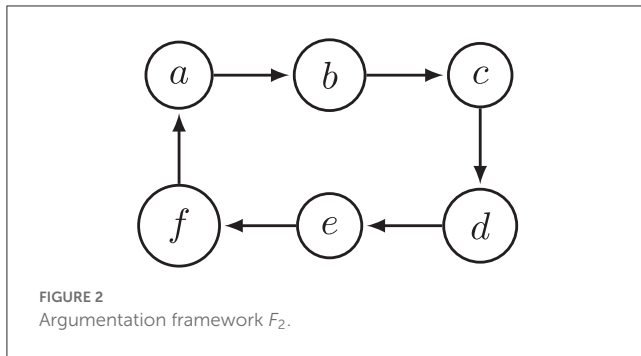
then $a \in A$.

DEFINITION 18. Let σ be an argumentation semantics. We say that σ satisfies *Strong Completeness Outside Odd Cycles (SCOOC)* iff for any AF F , every σ -extension of F is strongly complete outside odd cycles.

Before motivating the SCOOC principle, we first illustrate it with an example of a violation of the principle in the CF2 semantics.

EXAMPLE 2. Consider the argumentation framework F_2 depicted in Figure 2. It is a simple six-cycle. One of the CF2 extensions of F_2 is $E = \{a, d\}$. F_2 contains no odd cycle, so in particular b and c are not in an odd cycle. Since $\{c\}^- = \{b\}$, this means that no argument in $\{c\} \cup \{c\}^-$ is in an odd cycle. Moreover, $E \cap \{c\}^- = \emptyset$. Thus, for E to be strongly complete outside odd cycles, it would have to contain c . However, $c \notin E$, so E is not strongly complete outside odd cycles. We have, therefore, established that the CF2 semantics does not satisfy SCOOC.

The SCOOC principle is related to the property of *strong completeness*: An extension E is *strongly complete* iff every argument not attacked by E is in E . We call this property *strong completeness* as it is a strengthening of completeness, which states that every argument defended by E is in E .



The stable semantics is the only widely studied argumentation semantics that satisfies strong completeness. More precisely, the stable semantics can be characterized by the conjunction of conflict freeness and strong completeness. In other words, one can say that the stable semantics is motivated by the idea that a violation of strong completeness constitutes a paradox and should therefore be avoided.

The stable semantics satisfies strong completeness at the price of allowing for situations in which there are no extensions, and hence no judgment can be made on any argument whatsoever. Such cases are always due to odd *att*-cycles. So, we can say that odd *att*-cycles—unless resolved through arguments attacking the odd cycle—cause paradoxical situations. The idea of most semantics other than stable semantics is to somehow contain these paradoxes, so that they do not affect our ability to make judgments about completely or sufficiently unrelated arguments.

The idea of the SCOOOC principle is that while in odd cycles we may not be able to avoid paradoxical judgments about the arguments, i.e., a judgment in which an argument is not accepted even though none of its attackers is accepted, such paradoxical judgments should be completely avoided outside of odd cycles.

How does that differ from the containment of paradoxical situations provided by existing semantics? Admissibility-based semantics do not allow for any judgment about an argument in an unattacked odd cycle; however, this undecided status is not limited to odd cycles but carries forward to arguments that are not in an odd cycle but that are *att*-reachable from an odd cycle.

Naive-based semantics like CF2, stage, and stage2 allow for judgments about arguments in an unattacked odd cycle but also at the cost of affecting the way arguments that are not in odd cycles are interpreted. For example, as established in Example 2 earlier, CF2 allows for a six-cycle to be interpreted in a doubly paradoxical way despite the fact that it is an even cycle that can be interpreted in a non-paradoxical manner. This behavior of CF2 was also considered problematic by Dvořák and Gaggl (2016), who used this example to motivate their stage2 semantics, but as we will show in Figure 6, stage2 also fails to avoid paradoxical judgments about arguments that are not themselves involved in an odd cycle.

The SCOOOC principle was designed to systematically identify whether a semantics suffers from this problem. As it turns out, all the standard semantics other than stable do suffer from the problem, i.e., do not satisfy SCOOOC.

We will now look at which semantics satisfy or do not satisfy each of the two principles that we have defined.

THEOREM 1. The grounded, complete, naive, and *nsa*(CF2) semantics satisfy INRA.

Before we can prove the theorem, we first need some auxiliary definitions and lemmas.

DEFINITION 19. A semantics σ is called *SCC-rich* iff for every AF $F = \langle Ar, att \rangle$ such that $|SCCs(F)| = 1$ and every argument $a \in Ar$, there is an extension $E \in \sigma(F)$ such that E does not attack a .

DEFINITION 20. A semantics is called *semi-rich* iff for every AF $F = \langle Ar, att \rangle$ and every argument $a \in Ar$ such that $(a, a) \notin att$, there is an extension $E \in \sigma(F)$ such that E does not attack a .

DEFINITION 21. A semantics is called *SCC-semi-rich* iff for every AF $F = \langle Ar, att \rangle$ such that $|SCCs(F)| = 1$ and every argument $a \in Ar$ such that $(a, a) \notin att$, there is an extension $E \in \sigma(F)$ such that E does not attack a .

LEMMA 1. Naive semantics is semi-rich and thus also SCC-semi-rich.

PROOF. Let $F = \langle Ar, att \rangle$ be an AF and let $a \in Ar$ be an argument such that $(a, a) \notin att$. Let E be a naive extension of $F|_{Ar \setminus (\{a\} \cup \{a\}^-)}$. Then, $E \cup \{a\}$ is a naive extension of F and $E \cup \{a\}$ does not attack a . \square

LEMMA 2. Grounded and complete semantics are SCC-rich.

PROOF. Let $F = \langle Ar, att \rangle$ be an AF such that $|SCCs(F)| = 1$ and let $a \in Ar$. We distinguish two cases:

1. $att = \emptyset$. In this case, Ar is the only grounded and complete extension of F , and Ar does not attack a .
2. $att \neq \emptyset$. Since $|SCCs(F)| = 1$, this implies that every argument is attacked by some argument. Thus \emptyset is a grounded and complete extension of F . Since \emptyset does not attack a , the required condition is satisfied.

\square

The following lemma has a very technical proof that we provide in Appendix 1. Here, we just sketch the main idea of the proof and then discuss what is the main difficulty in making the argument rigorous.

LEMMA 3. Let σ be an SCC-rich or SCC-semi-rich semantics.

1. If σ is SCC-rich, then $scc(\sigma)$ satisfies INRA.
2. If σ is SCC-semi-rich, then $nsa(scc(\sigma))$ satisfies INRA.

PROOF SKETCH. First, we observe that for showing that $nsa(scc(\sigma))$ satisfies INRA, it is enough to consider AFs without self-attacking arguments. However, in such AFs, SCC-richness, and SCC-semi-richness coincide. So, we can actually assume SCC-richness for both parts of the lemma.

We consider an argument a that is attacked by every extension and need to show that removing that argument from the AF will not result in the emergence of new extensions or the disappearance of any previous extensions. Due to the SCC-richness of σ , a cannot be in an initial SCC. Instead, a must be in a position where, whatever happens in the SCCs that come before a , some argument attacking

a will be accepted. Thus, the SCC-recursive scheme removes a from the computation of the semantics at that step. Since that is the case, removing a from the AF will make no difference because what happens in the SCCs that preceded a will not be affected by the initial removal of a , and starting at the SCC that (originally) contains a , it makes no difference whether a is initially removed from the framework or removed from the computation by the SCC-recursive scheme due to having an attacker from a previous SCC.

□

The main difficulty in making this proof sketch a rigorous proof is that the removal of a may change the structure of the SCCs, as the SCC containing a may be split up into multiple SCCs. That complicates the argument significantly, but the rigorous proof in [Appendix 1](#) spells out in detail how the argument works to cover this case.

PROOF OF THEOREM 1. By Lemmas 1, 2, and 3 and the fact that $\text{grounded} = \text{scc}(\text{grounded})$, $\text{complete} = \text{scc}(\text{complete})$, and $\text{nsa}(\text{CF2}) = \text{nsa}(\text{scc}(\text{naive}))$, it directly follows that grounded , complete and $\text{nsa}(\text{CF2})$ satisfy INRA.

We now show that naive semantics satisfies INRA. Let $F = \langle Ar, att \rangle$ be an AF and let $a \in Ar$ be an argument such that for every $E \in \text{naive}(F)$, E attacks a . By the semi-richness of the naive semantics (Lemma 1), it follows that $(a, a) \in att$.

We need to show that $\text{naive}(F) = \text{naive}(F^{-a})$. Let $S \in \text{naive}(F)$. As $a \notin S$, $S \subseteq Ar \setminus \{a\}$. S is conflict free, and as S is maximal with this property in F , it is also maximal with this property in F^{-a} . So $S \in \text{naive}(F^{-a})$, as required.

Now, let $S \in \text{naive}(F^{-a})$. S is conflict free. Since $(a, a) \in att$, $S \cup \{a\}$ is not conflict free. Together with the maximality of S in F^{-a} , this implies that S is a maximally conflict free subset of Ar , i.e., $S \in \text{naive}(F)$, as required. □

THEOREM 2. Stable, preferred, semi-stable, stage, stage2, and CF2 semantics violate INRA.

PROOF. The fact that the preferred semantics violates INRA was already established in Example 1 with reference to the argumentation framework F_1 . The same argumentation framework also constitutes a violation of INRA for the stable, semi-stable, stage, and stage2 semantics, as these semantics coincide with the preferred semantics on F_1 and F_1^{-b} . A counterexample of CF2 semantics is shown in [Figure 3](#), as explained in the caption of the figure.

THEOREM 3. Stable semantics satisfies SCOOC.

PROOF. Consider an AF F , a stable extension E of F and an argument $a \in Ar$, such that $E \cap \{a\}^- = \emptyset$. Then, by definition of stable semantics, we have $a \in E$. Consequently, E is strongly complete, and in particular, E is strongly complete outside odd cycles.

THEOREM 4. Complete, grounded, preferred, semi-stable, naive, stage, CF2, stage2, and $\text{nsa}(\text{CF2})$ semantics violate SCOOC.

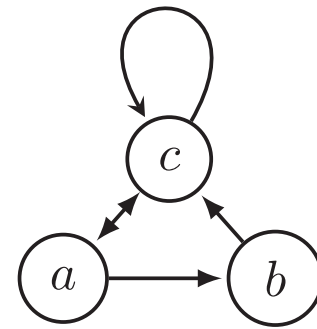


FIGURE 3

Argumentation framework F_3 . It shows that CF2 semantics violates INRA since both extensions $\{a\}$ and $\{b\}$ attack c , but after removing c , $\{b\}$ is no longer an extension.

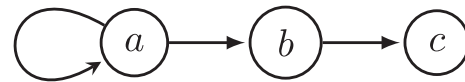


FIGURE 4

Argumentation framework F_4 . It shows that complete, grounded, preferred, and semi-stable semantics violate SCOOC since $E = \{a\}$ is an extension, but E is not strongly complete outside odd cycles: b and c are not in an odd cycle, $\{c\}^- = \{b\}$, but E does not contain c .

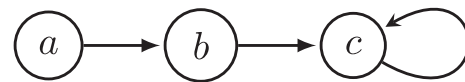


FIGURE 5

Argumentation framework F_5 . It shows that stage and naive semantics violate SCOOC since $E = \{b\}$ is an extension, but E is not strongly complete outside odd cycles: a is not in an odd cycle, $\{a\}^- = \{b\}$, but E does not contain a .

PROOF. The counterexample of CF2 was already presented in Example 2. The argumentation framework F_2 from that example (the simple six-cycle) also constitutes a counterexample of naive and $\text{nsa}(\text{CF2})$, as they agree with CF2 on the simple six-cycle.

A counterexample of complete, grounded, preferred, and semi-stable is shown in [Figure 4](#), and a counterexample of naive and stage is shown in [Figure 5](#), and a counterexample of stage2 is shown in [Figure 6](#).

Note that for a framework that does not contain any odd cycles at all, the preferred, and semi-stable extensions coincide with the stable extensions, so that in this special case, the SCOOC principle is also satisfied for the preferred and semi-stable semantics.

4. SCF2 semantics

In this section, we define and study the new semantics SCF2, which satisfies both of the new principles introduced in the previous section and the three principles defined in the preliminaries. Furthermore, we will motivate the design choices

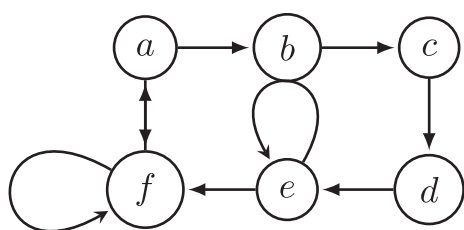


FIGURE 6

Argumentation framework F_6 . It shows that stage2 semantics violates SCOOC since $E = \{a, d\}$ is an extension, but E is not strongly complete outside odd cycles: b and c are not in an odd cycle, $\{c\}^- = \{b\}$, but E does not contain c .

in the definition of SCF2 by looking at how semantics defined in a similar way as SCF2 fail to satisfy at least one of directionality, INRA or SCOOC.

4.1. Definition of SCF2 and examples

We have seen in the previous section that $\text{nsa}(\text{CF2})$ satisfies INRA but does not satisfy SCOOC. The idea behind the definition of SCF2 is that we modify the definition of $\text{nsa}(\text{CF2})$ by already enforcing SCOOC at the level of the single SCCs considered in the SCC-recursive definition of $\text{nsa}(\text{CF2})$. For this, we define a variant of naive semantics called *SCOOC-naive semantics*.

DEFINITION 22. Let $F = \langle Ar, att \rangle$ be an AF, and let $A \subseteq Ar$. We say that A is an *SCOOC-naive extension* of F if A is subset-maximal among the conflict-free subsets of Ar that are strongly complete outside odd cycles.

Recall that CF2 is defined to be $\text{scc}(\text{naive})$, i.e., $\text{nsa}(\text{CF2}) = \text{nsa}(\text{scc}(\text{naive}))$. To define SCF2, we just replace naive semantics by SCOOC-naive semantics in this definition.

DEFINITION 23. SCF2 semantics is defined to be $\text{nsa}(\text{scc}(\text{SCOOC-naive}))$.

In other words, SCF2 works by first deleting all self-attacking arguments and then applying the SCC-recursive scheme that is also used in the definition of CF2, but applying SCOOC-naive semantics instead of naive semantics to each single SCC.

The computation of the SCF2 extensions of a given argumentation framework F can be described through the following non-deterministic algorithm:

1. Delete all self-attacking arguments from F .
2. Assign $E := \emptyset$.
3. Divide F into strongly connected components (SCCs).
4. Choose some initial SCC C of F .
5. Choose a maximal conflict-free subset A of C that satisfies the SCOOC principle.
6. Assign $E := E \cup A$.
7. Delete all arguments in C and all arguments attacked by A from F .
8. If F still contains arguments, go to step 3.

9. Return E .

EXAMPLE 3. Consider the argumentation framework F_7 depicted in Figure 7A. We describe how the four SCF2 extensions of F_7 can be computed using the above algorithm. First, we delete the self-attacking argument i . Then, we divide the resulting AF into SCCs as depicted in Figure 7B. The only initial SCC is $\{a, b, c, d, e, f\}$, so in step 4 of the algorithm, we choose C to be this SCC. Now in step 5, we have two choices:

- We can choose $A = \{b, d, f\}$. In this case, we delete arguments a, b, c, d, e, f , and j from F_7 . We return to step 3, and divide the AF into SCCs, as depicted in Figure 7C. There are two initial SCCs, $\{g\}$ and $\{k\}$. No matter which one we choose first, in the next step, we will have to choose A to be the completely chosen SCC. We then have one more iteration, in which we choose the set from $\{g\}$ and $\{k\}$ that we did not choose previously. Finally, the set E is $\{b, d, f, g, k\}$.
- We can choose $A = \{a, c, e\}$. In this case, we delete arguments a, b, c, d, e, f , and g from F_7 . We return to step 3, and divide the AF into SCCs, as depicted in Figure 7D. Now there are two initial SCCs, $\{h\}$ and $\{j, k, l\}$. Again, it does not matter in which order we choose them. Suppose we first choose h . Then, h gets added to E and deleted. In the final iteration, we need to choose the SCC $\{j, k, l\}$. Here, we can choose A to be $\{j\}$, $\{k\}$, or $\{l\}$. This gives rise to three possible values for the constructed extension, $\{a, c, e, h, j\}$, $\{a, c, e, h, k\}$, and $\{a, c, e, h, l\}$.

In order to allow readers to develop an intuition for how the SCF2 semantics behaves and how it differs from other semantics, we present in Table 1 the extensions of all example AFs considered in Section 3 according to the SCF2 semantics and all semantics introduced in Section 2.

4.2. Principle-based motivation for SCF2

As we will show below, SCF2 satisfies directionality, INRA, and SCOOC, which we have argued to be desirable principles when evaluating a semantics designed to correspond well to what humans would consider a rational judgment on the acceptability of arguments. The somewhat complex definition of SCF2 raises the question whether a simpler definition could also be enough to satisfy these three principles.

To approach this question systematically, we would like to point out that the definition of SCF2 contains three features that distinguish it from naive semantics: It starts by deleting all self-attacking arguments (the function nsa), it proceeds by applying the SCC-recursive scheme (the function scc), and within each SCC, it applies SCOOC-naive rather than naive semantics. If we consider each of these three features a switch that we can switch on or off, we have eight definitions of semantics, namely, naive, $\text{nsa}(\text{naive})$, SCOOC-naive, $\text{nsa}(\text{SCOOC-naive})$, $\text{scc}(\text{naive})$, $\text{nsa}(\text{scc}(\text{naive}))$, $\text{scc}(\text{SCOOC-naive})$, and $\text{nsa}(\text{scc}(\text{SCOOC-naive}))$. One can easily see that naive = $\text{nsa}(\text{naive})$, so these eight definitions define only seven different semantics, whose properties we now study in order to show that only SCF2 satisfies all three principles directionality, INRA, and SCOOC.

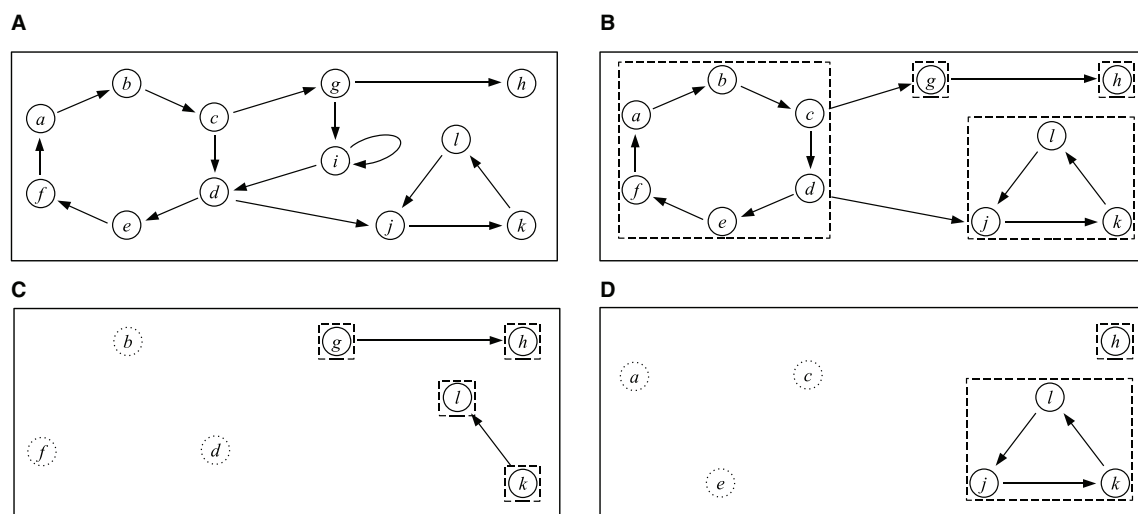


FIGURE 7

(A) Argumentation framework F_7 . (B–D) Intermediary steps in the computation of the SCF2 extensions of F_7 . Dashed lines indicate SCCs. Arguments in dotted circles have already been chosen to be included in the extension and are no longer part of the AF under consideration.

TABLE 1 Extensions of example AFs according to SCF2 and the semantics introduced in Section 2.

Semantics	F_1	F_2	F_3	F_4	F_5	F_6
SCF2	$\{a\}, \{b\}, \{c\}$	$\{a, c, e\}, \{b, d, f\}$	$\{a\}$	$\{b\}$	$\{a\}$	$\{a, c\}$
CF2	$\{a\}, \{b\}, \{c\}$	$\{a, d\}, \{b, e\}, \{c, f\},$ $\{a, c, e\}, \{b, d, f\}$	$\{a\}, \{b\}$	$\{b\}$	$\{a\}$	$\{a, c\}, \{a, d\},$ $\{b, d\}$
Naive	$\{a\}, \{b\}, \{c\}$	$\{a, d\}, \{b, e\}, \{c, f\},$ $\{a, c, e\}, \{b, d, f\}$	$\{a\}, \{b\}$	$\{b\}, \{c\}$	$\{a\}, \{b\}$	$\{a, c\}, \{a, d\},$ $\{b, d\}$
Stage2	$\{a\}$	$\{a, c, e\}, \{b, d, f\}$	$\{a\}$	$\{b\}$	$\{a\}$	$\{a, c\}, \{a, d\},$ $\{b, d\}$
Stage	$\{a\}$	$\{a, c, e\}, \{b, d, f\}$	$\{a\}$	$\{b\}$	$\{a\}, \{b\}$	$\{a, c\}, \{a, d\},$ $\{b, d\}$
Complete	$\emptyset, \{a\}$	$\emptyset, \{a, c, e\}, \{b, d, f\}$	$\emptyset, \{a\}$	\emptyset	$\{a\}$	\emptyset
Stable	$\{a\}$	$\{a, c, e\}, \{b, d, f\}$	$\{a\}$	—	—	—
Grounded	\emptyset	\emptyset	\emptyset	\emptyset	$\{a\}$	\emptyset
Preferred	$\{a\}$	$\{a, c, e\}, \{b, d, f\}$	$\{a\}$	\emptyset	$\{a\}$	\emptyset
Semi-stable	$\{a\}$	$\{a, c, e\}, \{b, d, f\}$	$\{a\}$	\emptyset	$\{a\}$	\emptyset

Furthermore, we also consider naivety and SCC-recursiveness, as these principles are important for getting a better picture of the behavior of SCF2 and allow us to conclude that SCF2 also satisfies several other principles studied in the literature, as we will discuss at the end of this section.

Table 2 shows which of these seven semantics satisfies which of these five principles (we use the standard name CF2 for $\text{scc}(\text{naive})$ and use the short name SCF2 to refer to $\text{nsa}(\text{scc}(\text{SCOOOC-naive}))$). Note that SCF2 satisfies all five principles, while no other of these seven semantics satisfies all five principles or even just the three principles directionality, INRA, and SCOOOC.

Thus, the complexity of the definition of SCF2 is not arbitrary but is required in the sense that all three differences between

the SCF2 semantics and the naive semantics (which has a much simpler definition) are needed to satisfy the considered principles. In other words, removing any non-empty subset of these three differences from the definition of the semantics would result in a semantics that does not satisfy all the considered principles.

We will now prove that every AF has an SCF2 extension and that the SCF2 semantics satisfies the five principles listed in Table 2. Concerning the other entries of Table 2, the results for CF2 and naive in the first three rows have been established in the literature (Baroni and Giacomin, 2007; van der Torre and Vesic, 2018), some of the results concerning INRA and SCOOOC have been shown in Section 3, and the remaining results are proven in Appendix 1.

TABLE 2 Properties of SCF2 and six semantics that are related to it with respect the five principles considered in this article.

	Naivety	Directionality	SCC-recursiveness	INRA	SCOOOC
naive = nsa(naive)	✓	×	×	✓	×
SCOOOC-naive	✓	×	×	×	✓
nsa(SCOOOC-naive)	✓	×	×	×	✓
CF2	✓	✓	✓	×	×
nsa(CF2)	✓	✓	✓	✓	×
scc(SCOOOC-naive)	✓	✓	✓	×	✓
SCF2	✓	✓	✓	✓	✓

First we need a lemma, whose rather long and technical proof can be found in [Appendix 1](#).

LEMMA 4. SCOOOC-naive semantics is SCC-semi-rich.

THEOREM 5. Every AF has at least one SCF2 extension.

PROOF. Lemma 4 implies that every single-SCC AF has a SCOOOC-naive extension. This, together with the definition of the SCC recursive scheme, implies that every AF has at least 1 s (SCOOOC-naive)-extension, and hence at least one SCF2 extension. \square

The proof of the following two theorems are in the appendix.

THEOREM 6. SCF2 satisfies naivety.

THEOREM 7. SCF2 satisfies directionality.

THEOREM 8. SCF2 satisfies SCC-recursiveness.

PROOF. From the definition of SCF2 it is immediately that $SCF2 = scc(SCF2)$ and that, therefore, SCF2 is SCC-recursive with base function $BF_5(F, C) := SCF2(F)$. \square

THEOREM 9. SCF2 satisfies SCOOOC.

PROOF. Consider an AF F , an SCF2 extension E of F , and an argument $a \in Ar$ such that no argument in $\{a\} \cup a^-$ is in an odd cycle and $E \cap a^- = \emptyset$. Then by definition of SCF2 semantics, the moment the SCOOOC-naive function is applied to a sub-framework of F containing a , we have $a \in E$. Consequently, E is strongly complete outside odd cycles. \square

THEOREM 10. SCF2 satisfies INRA.

PROOF. By Lemma 4, SCOOOC-naive semantics is SCC-semi-rich. So, by Lemma 3 and the definition of SCF2 it follows that SCF2 satisfies INRA. \square

Concerning the other principles studied in the literature, SCF2 has almost the same properties as CF2, the only exception being the succinctness principle ([van der Torre and Vesic, 2018](#)). This is proven in [Appendix 3](#). Most of the positive results follow directly from the results established above using logical relationships between principles that have been established in the literature ([van der Torre and Vesic, 2018](#)) – here, naivety and SCC-recursiveness play a crucial role.

5. Empirical cognitive studies

[Rahwan et al. \(2010\)](#) argued that artificial intelligence research will benefit from the interplay between logic and cognition and that; therefore, “logicians and computer scientists ought to give serious attention to cognitive plausibility when assessing formal models of reasoning, argumentation, and decision making.” Based on the observation that in the previous literature on formal argumentation theory, an example-based approach and a principle-based approach were used to motivate and validate argumentation semantics, they propose to complement these approaches by an *experiment-based approach* that takes into account empirical cognitive studies on how humans interpret and evaluate arguments. They made a first contribution to this new approach by presenting and discussing the results of two such studies that they conducted in order to test the cognitive plausibility of simple and floating reinstatement ([Rahwan et al., 2010](#)).

While the argumentation frameworks used in [Rahwan et al.’s](#) studies could not distinguish between preferred semantics and naive-based semantics like CF2, two more recent studies by [Cramer and Guillaume \(2018b, 2019\)](#) addressed this issue. Both of these studies made use of a group discussion methodology that is known to stimulate more rational thinking. According to the results of the first study ([Cramer and Guillaume, 2018b](#)), CF2, SCF2, stage, and stage2 semantics are significantly better predictors for human judgments on the acceptability of arguments than admissibility-based semantics like grounded, preferred, complete or semi-stable (all p -values < 0.001), but this study did not involve argumentation frameworks that allow distinguishing between CF2, SCF2, stage, and stage2 semantics.

According to the results of the second study ([Cramer and Guillaume, 2019](#)), SCF2, CF2, and grounded semantics are better predictors for human judgments on the acceptability of arguments than stage, stage2, preferred or semi-stable semantics (all $p < 0.001$). In addition, the results suggest that SCF2 is a better predictor than CF2 and grounded semantics, but the results are not significant.² We will now explain these results in more depth.

² While the SCF2 semantics had not yet been proposed at the time when this study and the two studies mentioned before were conducted, the design of the studies was such that they were not specifically tailored toward the semantics that the results were compared to in the articles about the studies. In other words, the results of the studies can be equally compared to any argumentation semantics whatsoever. Here, we compare them to the SCF2

As explained in Section 3, Dvořák and Gaggl (2016) critique a feature of CF2 semantics, namely, that in the case of a six-cycle, as depicted in Figure 2, CF2 allows two opposite arguments (e.g., *a* and *d*) to be accepted together. The second study by Cramer and Guillaume (2019) confirms that this criticism is in line with human judgments of argument acceptability. We briefly summarize the data on which this judgment is made (a more detailed explanation can be found in Cramer and Guillaume, 2019): Based on the overall responses of the participants in the study, Cramer and Guillaume pointed out that 12 of the 61 participants of their study have a high frequency of incoherent responses, so that they disconsider them from the further analysis. Among the remaining 49 participants, 22 follow a simple cognitive strategy of marking arguments as *Undecided* whenever there is a reason for doubt (in line with the grounded semantics), while 27 participants do not follow this strategy. Cramer and Guillaume called these 27 participants the *coherent non-grounded participants*.

In the case of 11 out of the 12 argumentation frameworks considered in the study, the majority of these 27 coherent non-grounded participants make judgments that are in line with CF2 semantics. The only exception to this is an argumentation framework involving a six-cycle, in which only 33% of the coherent non-grounded participants make a judgment in line with CF2 semantics, while 60% make a judgments that are similar in line with SCF2, stage2, preferred and semi-stable semantics.

Dvořák and Gaggl (2016) themselves had used this criticism against CF2 to motivate their stage2 semantics, but in the study by Cramer and Guillaume (2019), stage2 performed worse than CF2, as all other AFs in which stage2 and CF2 had different predictions were evaluated by most participants (including most coherent non-grounded participants) more in line with CF2 than with stage2.

In combination with the principle-based argument for SCF2 presented in the previous two sections, this provides additional support for our claim that SCF2 corresponds well to what humans consider a rational judgment on the acceptability of arguments.

6. Related work

The principle-based analysis of argumentation semantics was initiated by Baroni and Giacomin (2007) to choose among the many extension-based argumentation semantics that have been proposed in the formal argumentation literature. The handbook chapter of van der Torre and Vesic (2018) gives a classification of 15 alternatives for argumentation semantics using 27 principles discussed in the literature on abstract argumentation. Dvořák and Gaggl (2016) introduced stage2 semantics by showing how it satisfies various desirable properties, similar to how we motivate SCF2 semantics in this article.

Moreover, additional extension-based argumentation semantics and principles have been proposed by various authors. For example, Besnard et al. (2016) introduced a system for specifying semantics in abstract argumentation called SESAME. Moreover, many principles have been proposed for alternative semantics of argumentation frameworks, such as ranking

semantics (Amgoud and Ben-Naim, 2013), and for extended argumentation frameworks, for example, for abstract dialectical frameworks (Brewka et al., 2018).

The principle of Irrelevance of Necessarily Rejected Arguments is closely related to the well-studied area of dynamics of argumentation, in which also various principles have been proposed which are closely related to INRA. Cayrol et al. (2008) were maybe the first to study revision of frameworks using a principle-based analysis, and they have been related to notions of equivalence (Baumann, 2012; Oikarinen and Woltran, 2011). (Boella et al., 2009) defined principles for abstracting (i.e., removing) an argument, and (Rienstra et al., 2015) defined a variety of persistence and monotony properties for argumentation semantics. Our INRA principle is inspired by and closely related to the *skeptical IO monotony principle* they define. The difference is that their principle considers adding an attack rather than removing an argument.

After the INRA principle was proposed in the workshop article (Cramer and van der Torre, 2019) on which the current article is based, Cramer and Spörl (2021) studied the INRA principle in connection with the notion of admissibility and developed a new admissibility-based semantics—the *choice-preferred semantics*—that satisfies INRA.

The study of semantics and principles for abstract argumentation remains an active area of research. During the past few years, various new semantics have been proposed that are neither admissibility based nor naive based (Dvořák et al., 2022). These semantics were mainly motivated by the idea that self-attacking arguments should not affect the acceptance of other arguments, which has been called ambiguity blocking or undecidedness blocking. For these and other semantics, it remains to be checked whether they satisfy the INRA and SCOOC principles introduced in this article.

In addition to the cognitive studies on formal argumentation that are already mentioned in Section 5, several other such studies have been conducted. Cerutti et al. (2021) give an overview of empirical cognitive studies about formal argumentation. Concerning investigations into the relation between argumentation semantics from abstract argumentation on the one hand and human argument evaluation on the other, this overview article only lists the articles already mentioned in Section 5. The remaining articles mentioned in the overview article by Cerutti et al. are concerned with argumentation formalisms from other areas of formal argumentation like structured argumentation [e.g., Cerutti et al. (2014) and Yu et al. (2018)] as well as probabilistic and bipolar argumentation (e.g., Polberg and Hunter, 2018). Since these studies are about other areas of formal argumentation, they are not directly relevant to the research question addressed in this article. Concerning studies on abstract argumentation, there is also a recent article by Guillaume et al. (2022) that gives a more detailed analysis of the results from the study first presented in Cramer and Guillaume (2018b).

7. Conclusion and future work

Motivated by empirical cognitive studies on argumentation semantics, we have introduced a new naive-based argumentation

semantics in addition to the semantics already considered in the original articles about the studies.

semantics called SCF2. A principle-based analysis shows that it has two distinguishing features:

1. If an argument is attacked by all extensions, then it can never be used in a dialogue and, therefore, it has no effect on the acceptance of other arguments. We call it *Irrelevance of Necessarily Rejected Arguments*.
2. Within each extension, if none of the attackers of an argument is accepted and the argument is not involved in a paradoxical relation, then the argument is accepted. We define paradoxicality as being part of an odd cycle, and we call this principle *Strong Completeness Outside Odd Cycles*.

We have argued that these features, together with further satisfied principles and the findings from empirical cognitive studies, make SCF2 a good candidate for an argumentation semantics that corresponds well to what humans consider a rational judgment on the acceptability of arguments.

Though many results have been obtained—some of them listed in the appendix—there is also some work left to be done. First of all, for a few principles discussed in the literature, it still needs to be shown whether they hold for SCF2 or not. Moreover, dialogue-based decision procedures must be defined, and the complexity of the various decision problems must be established. Finally, an extension toward structured argumentation should be investigated.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Author contributions

While MC has done most of the work, LT helped with proving a significant amount of the formal results listed in [Appendix 3](#)

and with improving the wording and presentation of the whole paper. Both authors contributed to the article and approved the submitted version.

Funding

LT was financially supported by the Fonds National de la Recherche Luxembourg through the project Deontic Logic for Epistemic Rights (OPEN O20/14776480) and the Chist-Era grant CHIST-ERA19-XAI (G.A. INTER/CHIST/19/14589586), and by the European Union's Justice programme under grant agreement 101007420 (ADELE).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2023.1045663/full#supplementary-material>

References

- Amgoud, L., and Ben-Naim, J. (2013). "Ranking-based semantics for argumentation frameworks," in *Scalable Uncertainty Management*, eds W. Liu, V. S. Subrahmanian, and J. Wijsen (Berlin; Heidelberg: Springer Berlin Heidelberg), 134–147.
- Baroni, P., Caminada, M., and Giacomin, M. (2018). "Abstract argumentation frameworks and their semantics," in *Handbook of Formal Argumentation*, eds Baroni, P., Gabbay, D., Giacomin, M., and L. van der Torre (College Publications), 159–236.
- Baroni, P., and Giacomin, M. (2007). On principle-based evaluation of extension-based argumentation semantics. *Artif. Intell.* 171, 675–700. doi: 10.1016/j.artint.2007.04.004
- Baroni, P., Giacomin, M., and Guida, G. (2005). SCC-recursiveness: a general schema for argumentation semantics. *Artif. Intell.* 168, 162–210. doi: 10.1016/j.artint.2005.05.006
- Baumann, R. (2012). Normal and strong expansion equivalence for argumentation frameworks. *Artif. Intell.* 193, 18–44. doi: 10.1016/j.artint.2012.08.004
- Besnard, P., Doutre, S., Ho, V. H., and Longin, D. (2016). "SESAME-a system for specifying semantics in abstract argumentation," in *Proceedings of the First International Workshop on Systems and Algorithms for Formal Argumentation (SAFA) co-located with the 6th International Conference on Computational Models of Argument (COMMA 2016)*, eds M. Thimm, F. Cerutti, H. Strass, and M. Vallati, Potsdam, Germany, September 13, 2016, volume 1672 of CEUR Workshop Proceedings (Potsdam: CEUR-WS.org), 40–51.
- Boella, G., Kaci, S., and van der Torre, L. W. N. (2009). "Dynamics in argumentation with single extensions: abstraction principles and the grounded extension," in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 10th European Conference, ECSQARU 2009*, eds C. Sossai and G. Chemello, Verona, Italy, July 1–3, 2009. Proceedings, volume 5590 of Lecture Notes in Computer Science (Verona: Springer), 107–118.
- Brewka, G., Ellmauthaler, S., Strass, H., Wallner, J., and Woltran, S. (2018). *Abstract Dialectical Frameworks*. College Publications, International.
- Budzynska, K., and Villata, S. (2018). "Processing natural language argumentation," in *Handbook of Formal Argumentation*, eds P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre (College Publications), 577–627.
- Caminada, M. W. A., Carnielli, W. A., and Dunne, P. E. (2012). Semi-stable semantics. *J. Log. Comput.* 22, 1207–1254. doi: 10.1093/logcom/exr033
- Cayrol, C., de Saint-Cyr, F. D., and Lagasque-Schiex, M. (2008). "Revision of an argumentation system," in *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference, KR 2008*, eds G. Brewka and J. Lang (Sydney, NSW: AAAI Press), 124–134.
- Cerutti, F., Cramer, M., Guillaume, M., Hadoux, E., Hunter, A., and Polberg, S. (2021). "Empirical cognitive studies about formal argumentation," in *Handbook of Formal Argumentation, Vol. 2* (College Publications).

- Cerutti, F., Tintarev, N., and Oren, N. (2014). "Formal arguments, preferences, and natural language interfaces to humans: an empirical evaluation," in *Proceedings of the 21st ECAI 2014*, T. Schaub, G. Friedrich, and B. O'Sullivan, 207–212.
- Cramer, M., and Guillaume, M. (2018a). "Directionality of attacks in natural language argumentation," in *Proceedings of the Workshop on Bridging the Gap between Human and Automated Reasoning*, Vol. 2261, ed C. Schon (RWTH Aachen University, CEUR-WS.org), 40–46. Available online at: <http://ceur-ws.org/Vol-2261/>
- Cramer, M., and Guillaume, M. (2018b). "Empirical cognitive study on abstract argumentation semantics," in *Frontiers in Artificial Intelligence and Applications*, 413–424.
- Cramer, M., and Guillaume, M. (2019). "Empirical study on human evaluation of complex argumentation frameworks," in *Proceedings of JELIA 2019*. Available online at: http://icr.uni.lu/mcramer/downloads/2019_JELIA.pdf
- Cramer, M., and Spörl, Y. (2021). "The choice-preferred semantics for relevance-oriented acceptance of admissible sets of arguments," in *International Conference on Logic and Argumentation* (Springer), 94–111.
- Cramer, M., and van der Torre, L. (2019). "SCF2-an argumentation semantics for rational human judgments on argument acceptability," in *Proceedings of the 8th Workshop on Dynamics of Knowledge and Belief (DKB'19) and the 7th Workshop KI Kognition (KIK'19)*, volume 2445 of *CEUR Workshop Proceedings*, eds C. Beierle, M. Ragni, F. Stolzenburg, and M. Thimm (CEUR-WS.org), 24–35.
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77, 321–357. doi: 10.1016/0004-3702(94)00041-X
- Dvorák, W., and Gaggl, S. A. (2016). Stage semantics and the SCC-recursive schema for argumentation semantics. *J. Logic Comput.* 26, 1149–1202. doi: 10.1093/logcom/exu006
- Dvorák, W., Rienstra, T., van der Torre, L., and Woltran, S. (2022). "Non-admissibility in abstract argumentation," in *Computational Models of Argument - Proceedings of COMMA 2022, Cardiff, Wales, UK, 14-16 September 2022*, volume 353 of *Frontiers in Artificial Intelligence and Applications*, eds F. Toni, S. Polberg, R. Booth, M. Caminada, and H. Kido (Cardiff: IOS Press), 128–139.
- Guillaume, M., Cramer, M., van der Torre, L., and Schiltz, C. (2022). Reasoning on conflicting information: an empirical study of formal argumentation. *PLoS ONE* 17, e0273225. doi: 10.1371/journal.pone.0273225
- Lauscher, A., Wachsmuth, H., Gurevych, I., and Glavas, G. (2021). Scientia potentia est-on the role of knowledge in computational argumentation. *CoRR*, abs/2107.00281. doi: 10.48550/arXiv.2107.00281
- Lawrence, J., and Reed, C. (2020). Argument mining: a survey. *Comput. Linguist.* 45, 765–818. doi: 10.1162/coli_a_00364
- Oikarinen, E., and Woltran, S. (2011). Characterizing strong equivalence for argumentation frameworks. *Artif. Intell.* 175, 1985–2009. doi: 10.1016/j.artint.2011.06.003
- Polberg, S., and Hunter, A. (2018). Empirical evaluation of abstract argumentation: supporting the need for bipolar and probabilistic approaches. *Int. J. Approx. Reason.* 93, 487–543. doi: 10.1016/j.ijar.2017.11.009
- Prakken, H. (2018). "Historical overview of formal argumentation," in *Handbook of Formal Argumentation*, eds P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre (College Publications), 75–143.
- Rahwan, I., Madakkatel, M. I., Bonnefon, J.-F., Awan, R. N., and Abdallah, S. (2010). Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cogn. Sci.* 34, 1483–1502. doi: 10.1111/j.1551-6709.2010.01123.x
- Rahwan, I., and Simari, G. R. (2009). *Argumentation in Artificial Intelligence*, 1st Edn. Springer Publishing Company, Incorporated.
- Rienstra, T., Sakama, C., and van der Torre, L. W. N. (2015). "Persistence and monotony properties of argumentation semantics," in *Theory and Applications of Formal Argumentation-Revised Selected Papers*, volume 9524 of *Lecture Notes in Computer Science*, eds E. Black, S. Modgil, and N. Oren (Springer), 211–225.
- van der Torre, L., and Vesic, S. (2018). "The principle-based approach to abstract argumentation semantics," in *Handbook of Formal Argumentation*, eds P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre (College Publications), 3–73.
- Van Eemeren, F., and Verheij, B. (2018). "Argumentation theory in formal and computational perspective," in *Handbook of Formal Argumentation*, eds P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre (College Publications), 3–73.
- Verheij, B. (1996). "Two approaches to dialectical argumentation: admissible sets and argumentation stages," in *Proceedings of the biannual International Conference on Formal and Applied Practical Reasoning (FAPR) Workshop* (Universiteit), 357–368.
- Yu, Z., Xu, K., and Liao, B. (2018). Structured argumentation: restricted rebut vs. unrestricted rebut. *Stud. Logic* 11, 3–17.



OPEN ACCESS

EDITED BY

Emmanuelle Dietz,
Airbus, Germany

REVIEWED BY

Helena Lindgren,
Umeå University, Sweden
Antonio Rago,
Imperial College London, United Kingdom

*CORRESPONDENCE

Federico Castagna
✉ fCastagna@lincoln.ac.uk

SPECIALTY SECTION

This article was submitted to
Machine Learning and Artificial Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

RECEIVED 15 September 2022

ACCEPTED 20 February 2023

PUBLISHED 23 March 2023

CITATION

Castagna F, Garton A, McBurney P, Parsons S,
Sassoon I and Sklar EI (2023) EQRbot: A chatbot
delivering EQR argument-based explanations.
Front. Artif. Intell. 6:1045614.
doi: 10.3389/frai.2023.1045614

COPYRIGHT

© 2023 Castagna, Garton, McBurney, Parsons,
Sassoon and Sklar. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).
The use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

EQRbot: A chatbot delivering EQR argument-based explanations

Federico Castagna^{1*}, Alexandra Garton¹, Peter McBurney²,
Simon Parsons¹, Isabel Sassoon³ and Elizabeth I. Sklar⁴

¹School of Computer Science, University of Lincoln, Lincoln, United Kingdom, ²Department of Informatics, King's College London, London, United Kingdom, ³Department of Computer Science, Brunel University London, London, United Kingdom, ⁴Lincoln Institute for Agri-Food Technology, University of Lincoln, Lincoln, United Kingdom

Recent years have witnessed the rise of several new argumentation-based support systems, especially in the healthcare industry. In the medical sector, it is imperative that the exchange of information occurs in a clear and accurate way, and this has to be reflected in any employed virtual systems. Argument Schemes and their critical questions represent well-suited formal tools for modeling such information and exchanges since they provide detailed templates for explanations to be delivered. This paper details the EQR argument scheme and deploys it to generate explanations for patients' treatment advice using a chatbot (EQRbot). The EQR scheme (devised as a pattern of Explanation-Question-Response interactions between agents) comprises multiple premises that can be interrogated to disclose additional data. The resulting explanations, obtained as instances of the employed argumentation reasoning engine and the EQR template, will then feed the conversational agent that will exhaustively convey the requested information and answers to follow-on users' queries as personalized Telegram messages. Comparisons with a previous baseline and existing argumentation-based chatbots illustrate the improvements yielded by EQRbot against similar conversational agents.

KEYWORDS

argument schemes, computational argumentation, chatbot, explainability, decision-support systems, healthcare, XAI

1. Introduction

Artificial Intelligence constitutes a powerful means when deployed for assisting people in making well-informed decisions. Such assistance is delivered as a set of recommendations on which a human, who is interacting with the AI-based system, has the final word. In the healthcare sector, decision support systems (DSS) prove to be especially useful since they mostly present: time-saving virtual assistance for practitioners; help for patients in self-managing their health conditions; better documentation, retrieval and presentation of data (which, as stated in [Fairweather et al. \(2020\)](#), is still required to be reliable by showing that its provenance is non-repudiable); and, finally, a substantial cost saving due to the partial automation and optimization (while preferring cheaper, but still effective, treatment options) of the workflow ([Sutton et al., 2020](#)). Several DSS employ advanced machine learning algorithms as their main AI reasoning mechanism, although they do not seem to provide robust evidence of improved diagnostic performance in clinical environments ([Vasey et al., 2021](#)). Other DSS employ computational argumentation instead as their AI reasoning mechanism. Indeed, as highlighted by [Lindgren et al. \(2020\)](#), the handling of inconsistent and conflicting knowledge is a common feature in medical decision-making processes when the opinions of several medical experts are solicited with regard to specific cases. Arguments can reflect the opinion of a single practitioner, of a general/local medical

guideline or even represent the viewpoint of a patient concerning a particular treatment. As an example of argumentation-driven clinical DSS (henceforth cDSS), the authors of Kökciyan et al. (2021) model medical recommendations *via* meta-level arguments that makes it possible to determine the ground on which the object-level arguments are justified or preferred. The work of Cyras et al. (2018) moves, instead, toward the creation of a cDSS that employs the structured argumentation formalism of ABA⁺ (stemming from the Assumption-Based Argumentation framework originally described in Bondarenko et al., 1997) for automated reasoning with conflicting clinical guidelines, patients' information and preferences. Multiple studies have also been conducted in the field of cDSS considering patients suffering from multimorbidities (as in Oliveira et al., 2018 and Chapman et al., 2019). Although the results thus far achieved have mostly been positive, in Bilici et al. (2018) the authors emphasize the need for further investigations regarding considerations of shared decisions, patients' preferences and social contexts, and a broader range of drug interactions (including food-drug interactions). Argumentation-based cDSS have been devised also in this specific research area: the CONSULT project (outlined in papers such as Essers et al., 2018; Balatsoukas et al., 2019; Kökciyan et al., 2019) introduces a data-driven decision support tool to help patients with chronic conditions manage their multimorbidities in collaboration with their carers and the health care professionals who are looking after them.

The drive to overcome ethical issues involving AI-based systems, along with distrust from their users, constitutes the reason for the recent interest in the field of Explainable AI (XAI). The idea is that the trustworthiness of AIs can be improved by building more transparent and interpretable tools capable of: explaining what the system has done, what it is doing now and what it is going to do next while disclosing salient information during these processes (Bellotti and Edwards, 2001). Nevertheless, Vilone and Longo (2021) point out that there is no general consensus upon an unambiguous definition of explanations and their essential properties. Drawing from social sciences studies, Miller (2019) identifies specific features that could help characterize explanations, all of which converge around a single conclusion: explanations are *contextual*. Similarly, Bex and Walton (2016) consider explanations as speech acts, differentiated by context from other locutions, used to help *understand* something. More precisely, explanations are a transfer of understanding from one party to another, where understanding is intended as common knowledge" shared between those parties. That said, there still remain many active issues concerning XAI. In Gunning et al. (2019), the authors present a (non-exhaustive) list of these challenges, that includes topics such as: accuracy vs. interpretability, the use of abstractions to simplify explanations or prioritizing competencies over decisions. Another problem is related to the end-user who is meant to receive the explanation. Indeed, the explaineer might be an individual with a specific background. Taking into account the different knowledge and clarification needs of each target user group will ensure the generation of more compelling explanations. From this perspective, it is interesting to notice that the research presented in Antaki and Leudar (1992), and more recently in Cyras et al. (2021), propose an account of explanations that is primarily argumentative. Similarly, the survey of Vassiliades et al. (2021) concludes that

using argumentation to justify why an event started, or what led to a decision, can enhance explainability. These intuitions are also backed by McBurney and Parsons (2021), where it is suggested that AI systems should adopt an argumentation-based approach to explanations. The advocated approach points toward Douglas Walton's Argument schemes (AS), thoroughly discussed in Walton et al. (2008).

The paper is structured as follows. Starting from a brief introduction of the required background notions in Section 2, we will propose a new dialectical tool for delivering cDSS recommendations: the EQR scheme, its corresponding critical questions, and the role that such a model plays in providing explanation within the clinical setting (Section 3). Section 4 articulates its implementation in the context of the CONSULT system, whereas Section 5 describes the chatbot (EQRbot) and its internal architecture. The bot conveys information starting from an instantiated EQR scheme around which pivots any additional answer to follow-on users' questions. Finally, Sections 6 and 7 provide a discussion and conclusion, respectively.

1.1. Contributions

The research outlined in this paper presents several original contributions. Expanding on the previous work of Castagna et al. (2022) that sketched the novel EQR scheme, we are going to (1) provide a more detailed description of the EQR scheme. Such a formal structure emerges as an effective model for conveying practical and theoretical information yielded as consequences of a presumptive reasoning formalization involving acting upon an expert opinion. The EQR scheme herein proposed proves to be particularly suited in concentrating relevant knowledge within a single explanation. For this reason, we devise (2) an implementation in the form of a chatbot (EQRbot) integrated into the CONSULT system. This bot delivers tailored EQR-based recommendations to patients, helping them self-managing their conditions. These recommendations also embed an additional layer of information: the rationale behind the instantiated scheme acceptability (i.e., its evaluation according to the considered argumentation framework). Finally, the EQRbot main procedure draws from our third contribution: (3) an algorithm for computing and delivering explanations, of which we provide (4) a formal analysis of the performance.

2. Background

The following background covers a concise summary of computational argumentation, along with a short overview of how argument schemes (and their clinically specialized version) have been employed in the literature to deliver medical explanations. The introduced formal definitions and models will prove useful in the next sections.

2.1. Computational argumentation

Informal studies on argumentation are underpinned by a rich literary heritage, but it is only in the past decades that logic-based models of argumentation have been intensively investigated as core components of AI-driven and Multi-Agent Systems (Chesnevar et al., 2000; Bench-Capon and Dunne, 2007). The seminal work conducted in Dung (1995) has been the starting point for most of the recent interest and research in the field of abstract argumentation and its argumentative characterizations of non-monotonic inferences. Indeed, the main strength of his approach is the simple and intuitive use of arguments as a means to formalize non-monotonic reasoning while also showing how humans handle conflicting information in a dialectical way. In a nutshell, the idea is that correct reasoning is related to the admissibility of a statement: the argument is acceptable (i.e., justified) only if it is defended against any counter-arguments. The core notion of Dung's abstract approach revolves around the definition of an argumentation framework, that is a pair $AF = \langle AR, attacks \rangle$, where AR is a set of arguments, and 'attacks' is a binary relation on AR , i.e., $attacks \subseteq AR \times AR$, such that $attacks(X, Y)$ denotes the conflict existing between an argument X and its target Y . In the same paper, the author proposes also different semantics to capture alternative (skeptical or credulous) types of reasoning:

Definition 1 (Argumentation semantics). Let $AF = \langle AR, attacks \rangle$, and $S \subseteq AR$ be a set of arguments:

- S is *conflict free* iff $\forall X, Y \in S: \neg attacks(X, Y)$;
- $X \in AR$ is *acceptable* w.r.t. S iff $\forall Y \in AR$ such that $attacks(Y, X): \exists Z \in S$ such that $attacks(Z, Y)$;
- S is an *admissible* extension iff $X \in S$ implies X is acceptable w.r.t. S ;
- An admissible extension S is a *complete* extension iff $\forall X \in AR: X$ is acceptable w.r.t. S implies $X \in S$;
- The least complete extension (with respect to set inclusion) is called the *grounded* extension;
- A maximal complete extension (with respect to set inclusion) is called a *preferred* extension.

As anticipated, AFs represent general frameworks capable of providing argumentative characterizations of non-monotonic logics.¹ That is to say, given a set of formulae Δ of some logical language L , AFs can be instantiated by such formulae. The conclusions of justified arguments defined by the instantiating Δ are equivalent to those obtained from Δ by the inference relation of the logic L . These instantiations paved the way for a plethora of different studies concerning the so-called "structured" argumentation (as opposed to the abstract approach). Among these, Besnard and Hunter (2008), Modgil and Prakken (2013), and Toni (2014) describe a formalization of arguments that follows the same model of the Argument Schemes introduced in Walton et al. (2008). That is to say, arguments are typically used to advocate a claim

based on the premises put forward as evidence to support such a claim.

2.2. Argument schemes and explanations in clinical settings

Argument schemes have been extensively investigated and employed in the AI literature as a way to directly convey presumptive reasoning in multi-agent interactions (for example, Atkinson et al., 2006; Tolchinsky et al., 2012; Grando et al., 2013). Each AS is characterized by a unique set of critical questions (CQs), rendered as attacking arguments, whose purpose is to establish the validity of the scheme instantiations. This generates an argumentation framework that can then be evaluated according to one of the semantics described in Dung (1995). Such evaluation embeds the rationale for choosing an argument over another, meaning that justified instantiations of schemes can be employed for conveying explanations. The use of argument schemes for providing explanations is, indeed, not unusual, especially in the clinical setting. In Shaheen et al. (2021), the authors introduce the *Explain Argument Scheme*, which models explanations based on the reasons, types (of reasons) and levels (of abstraction) and shows a (pro or con) rationale for giving a particular drug to a patient. The work presented in Sassoon et al. (2019), Kökciyan et al. (2020), and Sassoon et al. (2021) harnesses *Explanation Templates* that differ according to the reasoning and argument scheme represented and include placeholders for the actual instantiated variables specific to a given application of the scheme. Formally:

Definition 2 (Argument Scheme). $AS = \langle Prem, Con, Var \rangle$ denotes an argument scheme, where **Prem** is a set of premises, **Con** is the conclusion, and **Var** is the set of variables used in the argument scheme.

Definition 3 (Explanation Template). Let AS be an argument scheme (as per Definition 2), and txt be a natural language text that includes elements from Var . Then, an *Explanation Template* for AS can be rendered as the tuple $Expl_{AS} = \langle AS, txt \rangle$.

Definition 4 (Explanation). An explanation is a tuple $\langle Expl_{AS}, AS_i \rangle$ such that $Expl_{AS}$ is the explanation template introduced in Definition 3, AS_i is an acceptable (as per Definition 1) instantiation of AS with respect to some AF , and every variable in txt of $Expl_{AS}$ is instantiated by the corresponding element in AS_i .

Intuitively, Explanation Templates are engineered to be adaptive toward the circumstance of their employment and thus generate tailored explanations. That is to say, argument schemes model stereotypical patterns of reasoning in different generic situations, increasing their versatility of usage thanks to a number of integrated variables. Leveraging those variables, Definition 3 depicts formal structures that further enhance their flexibility by considering specific natural language snippets concerning the current context. These structures account then for explanations that enjoy the *contextuality* property (one of the most relevant features of explanations according to Miller, 2019), while they also acknowledge the end-users' different knowledge, understanding capability, and clarification needs.

¹ In Dung (1995), the author employs Reiter's Default logic (Reiter, 1980) and Pollock's Inductive Defeasible logic (Pollock, 1987) as an example of non-monotonic reasoning rendered via abstract argumentation.

2.3. Clinically specialized argument schemes

In order for a cDSS to provide the appropriate medical suggestions, explanation templates have previously been mapped to the *Argument Scheme for Proposed Treatment* (ASPT) (Sassoon et al., 2019, 2021; Kökciyan et al., 2020). Introduced in Kokciyan et al. (2018), ASPT derives from the *Argument Scheme for Practical Reasoning* as presented in Atkinson and Bench-Capon (2007). It instantiates an argument in support of a possible treatment, given the facts *Ft* about the patients and the goal *G* to be achieved.

ASPT
<div>Premise : Given the patient's fact <i>Ft</i> Premise : In order to realize goal <i>G</i> Premise : Treatment <i>T</i> promotes goal <i>G</i></div> <div>Conclusion : Treatment <i>T</i> should be considered</div>

As with each argument scheme, ASPT is accompanied by a series of critical questions that serve to assess the efficacy of the proposed treatment. In Sassoon et al. (2021), some of these questions are modeled as clinical specializations of existing argument schemes (listed in Walton et al., 2008) and cover particular aspects of the suggested treatment, such as *AS from Patient Medical History*, *AS from Negative Side Effect* and *AS for Contraindications*.

3. Methods: Providing explanations via the EQR argument scheme

3.1. EQR argument scheme

Devised as a model of Explanation-Question-Response agents interactions sketched in McBurney and Parsons (2021), the EQR argument scheme draws from the *AS for Practical Reasoning* (the variation of the AS presented in Walton (1996) as characterized in Atkinson and Bench-Capon, 2007) and the *AS from Expert Opinion* (Walton, 1997). The underlying idea is to merge the knowledge elicited by those two formal patterns in a single scheme that would then yield the advantage of concentrating and synthesizing the same amount of information in a unique data structure that may be queried more conveniently. That is to say, the purpose of the EQR scheme is to formalize the consequences arising (and the presumptive reasoning leading to them) by acting upon a specific expert opinion. A reference to such authority provides the rationale that justifies the conclusion of the argument, also leaving chances of inquiry for more detailed explanations.

The proposed scheme assumes the existence of:

- A finite set of knowledgeable experts, called *Experts*, denoted with elements *E*, *E'*, etc. Experts are deemed knowledgeable if they can somehow prove their competencies (e.g., years of experience, professional achievements, research publications).

EQR
<div>Premise : In the current state <i>R</i> Premise : acting upon α (from an expert <i>E</i> in a field <i>F</i>) Premise : will result in a new state <i>S</i> Premise : which will make proposition <i>A</i> true (alternatively, false) Premise : which will promote some value <i>v</i></div> <div>Conclusion : Acting upon the opinion α should make proposition <i>A</i> true (false) and entail value <i>v</i></div>

- A finite set of disciplinary fields of expertise, called *Fields*, denoted with elements *F*, *F'*, etc.
- A finite set of propositions, called *Opinions*, denoted with elements α , β , etc. Each member represents the viewpoint of an expert with regard to a specific topic.
- A finite set of propositions, called *Prop*, denoted with elements *A*, *B*, etc.
- A finite set of states, called *States*, denoted with elements *R*, *S*, etc. Every member describes a specific state of the world and corresponds to an assignment of truth values {**Truth**, **False**} to every element of *Prop*.
- A finite set of *Values* denoted with elements *v*, *w*, etc. This category includes both positive (i.e., constructive, such as wellbeing, altruism, integrity, etc.) and negative (i.e., non-constructive, such as dishonesty, manipulation, greed, etc.) values.
- A function *acting_upon* that maps each element of *Opinions* to a member of *States*.

Intuitively, starting from the current circumstance *R* and acting upon the opinion asserted by a competent expert in the relevant field, the agent instantiating the scheme wishes to attain *A* (or not *A*) and the actual reason for it (value *v*), along with the entailed consequences, whether they are desired or not (new state *S*). As an example of expert opinion, consider an architect asserting that, according to her recent evaluation, the nearby bridge requires immediate maintenance to prevent its collapse. In this case, by acting upon such an opinion, the practical intervention of specialized workers will change the state of the world into a new state where the bridge is no longer precarious (promoting the safety value).

The EQR scheme is accompanied by specifically designed critical questions:

- (EQR.CQ1) Is *E* the most knowledgeable expert source?
- (EQR.CQ2) Is *E* trustworthy?
- (EQR.CQ3) Is *E* an expert in the field *F* that α is in?
- (EQR.CQ4) Would acting upon α imply *A* (or not *A*)?
- (EQR.CQ5) Are there alternative experts' opinions that can be acted upon to imply *A* (or not *A*)?
- (EQR.CQ6) Would acting upon α entail contradictory propositions?
- (EQR.CQ7) Is *A* consistent with what other experts assert?
- (EQR.CQ8) Is α based on the (facts expressed by) state *R*?
- (EQR.CQ9) Is *F* the most relevant disciplinary field to *A* given the (facts expressed by) state *R*?
- (EQR.CQ10) Would acting upon α promote a negative value?

Following an approach akin to [Sassoon et al. \(2021\)](#), we can model each of the above critical questions into corresponding argument schemes. Each of these additional argument schemes may have its respective critical questions. However, we are omitting them since a full list of CQs for every possible argument scheme elicited by the critical questions of EQR is out of the scope of the current paper. For simplicity, we are going to outline only three of such templates.

3.1.1. AS for expert reliability (ASEXP)

ASEXP
<i>Premise</i> : Given a set of knowledgeable experts <i>Premise</i> : E is more trustworthy and knowledgeable than any other experts
<i>Conclusion</i> : E should be considered the most reliable expert

The *AS for Expert Reliability* fleshes out why a proficient source should be regarded as the most reliable (i.e., the most knowledgeable and trustworthy) in a group of several experts (if any). This is connected with and models EQR.CQ1-CQ2. Notice that here we are assuming a hierarchy of experts based on their reliability achieved by a preliminary probing of the ASEXP scheme instantiation (through its respective CQs) and the available professionals in the set of *Experts* that informs the EQR scheme instantiation. As an example, we could envisage a team of archaeologists at different stages of their careers. Everyone is considered an expert with several years of experience in their competence area. However, among them, there is a person (E) who has published more research articles and has participated in more archaeological excavations than any other member of the examined group of professionals (most knowledgeable). In addition, E has also diligently conducted the role of treasurer in each past expedition he took part in (trustworthy). Therefore, E can be deemed as the most reliable expert within those present. Observe that the same result will also occur if E is the only element of the considered set. Anticipating our implementation of the scheme within the CONSULT cDSS, let us also present another example that considers, like the aforementioned system, only clinical guidelines as Experts. This may yield an ASEXP instantiation where the World Health Organization (WHO) and other local practices are compared. WHO guidelines² (E), informed by several global professionals in a multitude of medical areas, result in the most knowledgeable source of expertise if measured against any other guidances based upon the proficiency of smaller (often not international) local practitioners teams, as occurs for hospital guidelines. The formers also emerge as the most trustworthy guidances since they are regularly inspected by a specific review committee composed of appropriately trained staff members. As such, E can be regarded as the most reliable expert among those present.

2 <https://www.who.int/publications/who-guidelines>

3.1.2. AS for relevant field of expertise (ASF)

ASF
<i>Premise</i> : Given a set of disciplinary fields of expertise <i>Premise</i> : Given the current state R <i>Premise</i> : Given a goal to achieve G <i>Premise</i> : F yields more connections, with respect to R and G, than any other fields
<i>Conclusion</i> : F should be considered the most relevant disciplinary field

The *AS for Relevant Field* provides the rationale for identifying the most relevant field, with respect to the current state of affairs R and a goal to achieve G, among a set of different disciplinary fields of expertise. This AS is correlated with and models EQR.CQ9. Once again, we are assuming a hierarchy of fields of expertise, based on their relevance over R and G, achieved by a preliminary probing of the ASF scheme instantiation (through its respective CQs) and the available elements in the set of *Fields* that informs the EQR scheme instantiation. As an example, consider R to be a state where a pandemic has spread to a whole country. To deal with such an emergency and promote people’s health (G), we should probably resort to epidemiology as a more relevant field of expertise rather than, say, oncology or neurology. That is because the former can be deemed as having more connections with R and G, hence proving to be more relevant than the latter.

3.1.3. AS for alternatives options (ASO)

ASO
<i>Premise</i> : Given a set of alternative options <i>Premise</i> : Given circumstance C <i>Premise</i> : Option O does not cause complications in circumstance C
<i>Conclusion</i> : O should be selected

The *AS for alternative Options* examines the reasons why a specific option, given a particular circumstance C, should be selected among a set of alternative options. This AS is correlated with and models EQR.CQ5. As an example, we can picture a man that needs to testify in court for a robbery he witnessed. Unfortunately, he also knows the thief. The man is now required to choose between producing a deposition that will incriminate his acquaintance or lying about having witnessed the crime at all. However, since perjury is a prosecutable criminal offense, telling the truth proves to be the only option that does not cause legal complications. As such, the witness will select the former alternative.

3.2. EQR and explanations in medical setting

Intuitively, the EQR scheme can display a large number of information bits to an *explainee* when looking for clarifications

about a proposed treatment. Notice indeed that the EQR scheme can encompass ASPT such that it renders: (i) the treatment T as the expert's opinion α (from an expert E in a field F); (ii) the patient fact Ft as part of the current state R and (iii) the goal to be realized G as proposition A . That is to say, by embedding ASPT into the EQR scheme, it will be possible to give more opportunities for inquiry to an agent seeking clinical recommendations. Certainly, in this way, further aspects can be interrogated and this can lead to more satisfactory (and complete) explanations. For example, the additional data comprised in the current state R , the connected field of expertise F , the immediate consequence S entailed by the proposed treatment, or the value v conveyed by the truth-value of A , all of these are elements that can be interrogated by the patients. In particular, knowing the source of the recommendation E (in the remainder of the paper, this will correspond to the chosen clinical guideline) may boost the patient's trust in the explainer and the advised medical care plan. Moreover, the rationale behind the provided explanations can be further investigated (resulting in additional, more detailed, explanations) thanks to the extra information supplied by the answers to each critical question and corresponding argument that informs valid instantiations of the EQR scheme (and the incorporated ASPT). This entails that the same CQs that challenges ASPT will also question instantiations of the EQR scheme when deployed for medical recommendations. For example, the CQs concerning the presence of contraindications and negative side effects within the proposed treatment (that structure *AS for Contraindications* and *AS from Negative Side Effect* and in the work of [Sassoon et al., 2021](#)) will revise the previously introduced *AS for alternative Options* in a clinically specialized form. The resulting *AS for alternative Clinical Options* (ASCO) describes the reasoning pattern that elicits the choice of a specific harmless treatment for a patient, considering her health conditions. Indeed, the selection of the recommended remedy is informed by the subject's health record: it thus strictly avoids any potentially dangerous medication. As an example, depict R as the state that includes a patient suffering from a bacterial chest infection. There are three available antibiotics that can treat such a disease in the current state R : amoxicillin³, cefalexin⁴, and azithromycin⁵. According to the information documented by the subject's medical facts (Ft) embedded in R , the patient is particularly sensitive to joint and muscle pain, which is listed among the amoxicillin side effects. Furthermore, azithromycin should be avoided due to its contraindications for people affected by heart problems, as, suppose, is our virtual subject. On the other hand, cefalexin (T) has already been administered to the patient in the past without resulting in any dangers or complications. As such, the latter is the treatment that should be recommended to cure the infection.

An EQR Explanation Template is then determined as in Definition 3, although it employs the EQR scheme rather than a generic AS. Similarly, we can formalize an instance of such a template as:

Definition 5 (EQR Explanation). An EQR explanation is a tuple $\langle \text{Expl}_{\text{EQR}}, \text{EQR}_i \rangle$ such that Expl_{EQR} is the explanation template

ASCO

Premise : Given a set of alternative treatments
 Premise : Given the current state R
 Premise : Considering the patient's fact Ft (subsumed in R),
 treatment T does not cause contraindication nor side effects

Conclusion : T should be recommended

for the EQR scheme, EQR_i is an acceptable (as per Definition 1) instantiation of the EQR scheme with respect to some AF_i and every variable in txt of Expl_{EQR} is instantiated by the corresponding element in EQR_i .

Example 1. Suppose that we have an acceptable (as per Definition 1) clinical instantiation of the EQR scheme, informed by its critical questions and a specific knowledge base. Assume also that the scheme variables $\text{Var} = \{R, E, F, \alpha, S, A, v\}$ are equivalent to the following:

$[R]$: the patient's previous health record and the current fever and headache (due to COVID-19)
 $[E]$: the NICE guidelines⁶
 $[F]$: medical management of COVID-19
 $[\alpha]$: the administering of paracetamol
 $[S]$: the reduction of fever and headache
 $[A]$: controlling the negative effect of the COVID-19 virus
 $[v]$: the patient's wellbeing

Finally, let txt be the natural language text: *Given $[R]$, the expertise of $[E]$ in the field of $[F]$ indicates $[\alpha]$ as an effective treatment. This should lead to $[S]$ which will bolster the goal of $[A]$ and promote $[v]$* . Then, the actual EQR Explanation would be:

"Given the patient's previous health record and the current fever and headache (due to COVID-19), the expertise of the NICE guidelines in the field of medical management of COVID-19 indicates the administering of paracetamol as an effective treatment. This should lead to the reduction of fever and headache which will bolster the goal of controlling the negative effect of the COVID-19 virus and promote the patient's wellbeing".

4. The CONSULT system

The CONSULT⁷ system is a novel data-driven mobile cDSS designed to help patients self-managing their condition and adhere to agreed-upon treatment plans in collaboration with healthcare professionals. Its main components are outlined in the following paragraphs and depicted in [Figure 1](#). More details on the architecture of the system are available in [Chapman et al. \(2022\)](#).

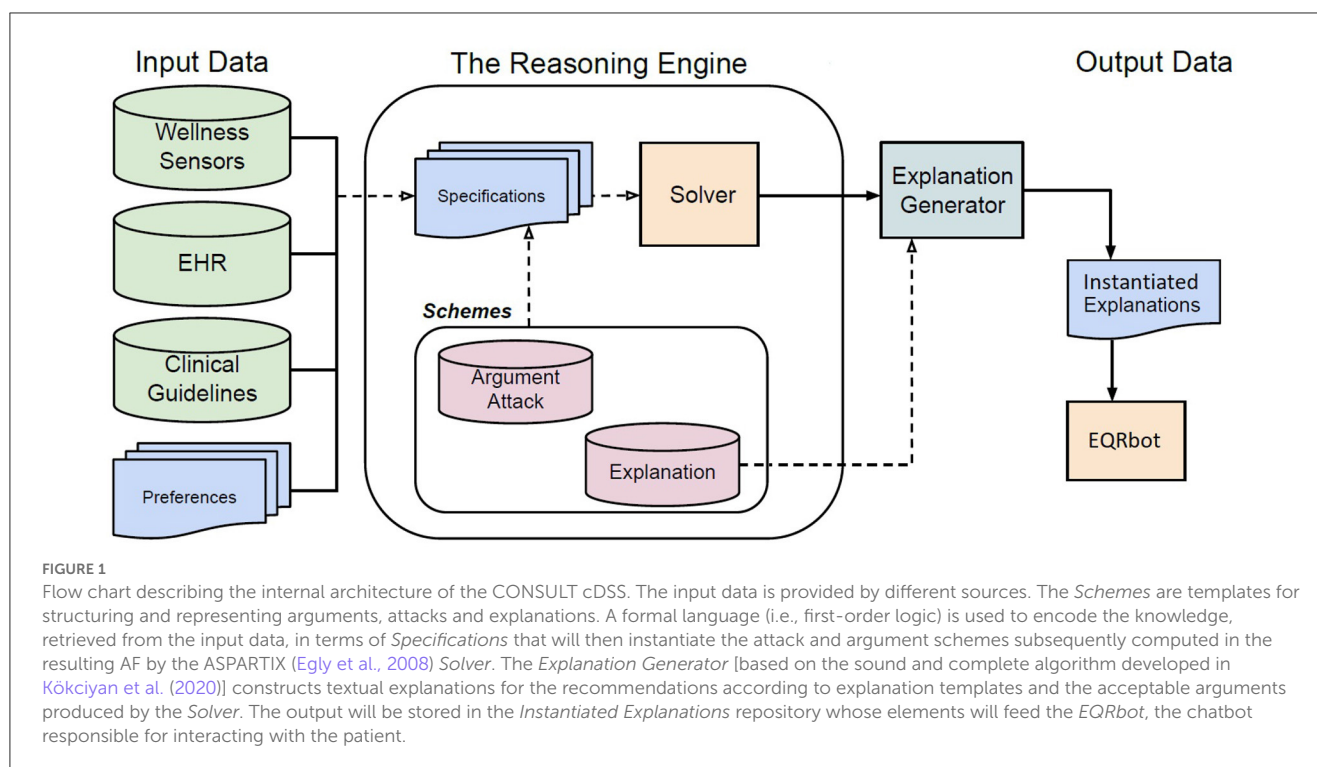
³ <https://www.nhs.uk/medicines/amoxicillin/>

⁴ <https://www.nhs.uk/medicines/cefalexin/>

⁵ <https://www.nhs.uk/medicines/azithromycin/>

⁶ <https://www.nice.org.uk/guidance>

⁷ <https://consultproject.co.uk>



4.1. Data inputs

There are three main types of data inputs into the CONSULT system: Wellness Sensors, Electronic Health Records and clinical guidelines. The *Wellness sensors* used included a Heart Rate monitor, a Blood Pressure Cuff and an ECG (Electrocardiogram) patch. The live parameters from these sensors are collected and displayed in one dashboard in the CONSULT system. This data is also used within the Argumentation Schemes instantiated in the reasoning engine. Information is additionally collected from the *Electronic Health Record (EHR)*, for example the patients' known allergies and prescriptions along with their general medical history. Finally the *clinical guidelines*, i.e., official documents published by medical organizations (as the already mentioned NICE guidelines), are also represented within the system. The CONSULT system also considers the *preferences* of stakeholders allowing for personalized recommendations. Such preferences are rendered as hierarchies of information (e.g., values, treatment, and guidelines) elicited from multiple sources, e.g., patient and treating clinician (which also convey the interests of the healthcare organization and the ethical oath they have to observe). Ultimately, tailored algorithms are used to map these medical data and preferences into the formal language used by the reasoning engine. That is to say, stored in a knowledge base (i.e., the *Specifications*), data is represented in terms of facts and Answer Set Programming (ASP) rules using first-order logic.

4.2. Specifications

The EHR data provides information such as the *current_state* of a patient (including demographics and current medications),

which need to be taken into account, along with the health parameters detected by the wellness sensors, when suggesting a treatment. Indeed, there may be age or other conditional restrictions related to the recommendation of, say, certain over-the-counter medications. For example, consider Frida, a pregnant patient currently suffering from fever and headache due to the COVID-19 virus. These facts will be formalized in first-order logic by the cDSS as *current_state*(fever, headache, COVID19) and *condition*(pregnancy). A treatment may then be recommended (as shown in Example 1) following the clinical guidelines of NICE-NG191⁸ and NHS⁹ (after their encoding into ASP-rules) that specifically handle those circumstances.

4.3. Schemes

Argument, attack and explanation schemes are templates representing common patterns of reasoning and relate a set of premises to a conclusion, all of which are sentences that can be represented in first-order logic and include variables that can be instantiated by data stored in a knowledge base. These schemes are kept in the *Schemes* repository and are rendered as ASP rules composed of a *rule body*, namely a conjunction of predicates (premises of the scheme), and a *rule head*, namely the scheme conclusion. The information stored in the *Specifications* data will

⁸ NICEcovid-managementguide section that specifically covers fever-management.

⁹ NHSwebpage section that specifically deals with ibuprofen assumption risks during pregnancy (redirected from NICE webpage).

then instantiate the elements of *Schemes* (i.e., attack and argument schemes) and thereupon will be fed to the *Solver*.

4.4. Solver and explanation generator

The argumentation-based reasoning engine runs on ASPARTIX (Egly et al., 2008), an ASP-Solver capable of computing arguments extensions under the required semantics (Dung, 1995). The reasoning engine leverages a formal representation of arguments through their respective argument schemes, critical questions and attacks to account for the conflicts between arguments in a given domain. The engine relies on the EvalAF algorithm to construct an argumentation framework for decision support and the ExpAF algorithm to provide explanations for acceptable arguments and attacks through the use of explanation templates¹⁰. The EvalAF algorithm generates an argumentation framework from a knowledge base and computes extensions under given semantics. The ExpAF algorithm maps acceptable arguments and attacks into explanations in natural language, using the sets of acceptable arguments and attacks, and corresponding explanation templates (Definition 3). In charge of the generation of such explanations is the sound and complete algorithm developed and implemented in Kökciyan et al. (2020).

4.5. Instantiated explanations

The *Instantiated Explanations* repository contains the rationales that justify the EQR explanation(s) (also member(s) of the repository) that serves as the pivotal element upon which all the other information is connected. Any answer to the questions moved by users of the CONSULT cDSS will be drawn from the data stored in such an archive. Notice that each explanation is tailored to the specific interacting patient's requirements, preferences and medical records. That is because the system manages only known information about the user and their conditions, thus providing suited routine recommendations conveniently retrieved by the applicable clinical guidelines (according to the predetermined cDSS resources and the patient's preferences). The user is made aware that CONSULT is not conceived to solve conflicts or handle unfamiliar data that would require professional medical expertise. Given this constraint, we can understand how the explanations stored within the *Instantiated Explanations* repository have to be finite.

5. EQRbot

The agent that will handle the interaction with the patient is a retrieval-type chatbot, i.e., a kind of bot that focuses on retrieving contexts and keywords from the user's prompts in order to select the best response to give.¹¹ The explanation process will occur as delineated in Figure 2. After having provided the initial explanation (i.e., the EQR explanation informed by an acceptable

instantiation of the EQR scheme), the patient will be asked to express their opinion. If the user is satisfied with the explanation, then the conversation will immediately end. Alternatively, the chatbot will demand: a brief context (e.g., "Would you please specify the context of your explanation request?") along with the actual request from the patient. Consider that the interaction is not limited by a specific set of options to which the explainee needs to comply: the choice of words to use for formulating the inquiries is completely unrestricted. By matching stored explanations (all of which account for the stakeholders' preferences), context and user input, the bot will output the additional solicited information. Observe that the double query prompted by the conversational agent ensures a significant reduction of misunderstandings when providing answers to the patient. That is because the matching occurs *via* a double-layer word similarity counter function based on a BoW (Bag of Words) model. The explainer (chatbot) can be considered successful in its clarification attempt if the proposed explanation is deemed satisfactory by the user. Recall that the patient is aware of the EQRbot's inability to address questions regarding information not stored within the CONSULT system. As such, a satisfactory explanation may also be depicted as the realization that the user has to contact an healthcare professional should they have further queries. This will stop the loop of answers/questions and will end the conversation. It will continue otherwise.

It should be noted that the presence of multiple initial acceptable EQR explanations will not affect the chatbot operations. Since all of the explanations are acceptable, there is no need to further invoke the reasoning engine. The explanations are all considered equally good, seeing that our criteria for presenting an explanation is its acceptability (in turn influenced by the stakeholders' preferences), and so the EQRbot will randomly choose one of the available options and will then begin its interaction with the user. To this end, observe also that the bot is designed to avoid any unnecessary prolongation of the interaction to focus only on the required explanations. For this reason, the EQRbot will not start a conversation (nor even send a message) without the user prompt, but will react to each received text.

5.1. NLP filter

The chatbot employs a Natural Language Processing (NLP) filter in order to refine the input it receives from the patient and the stored instantiated explanations (Figure 2). The filtering process comprises: (a) the separation of the considered data into lists of single words (tokenization); (b) the elimination of the most common English words, including conjunctions and prepositions (stop-words removal); (c) the transformation of each word into its lemmatic form (lemmatization). The purpose of this refinement procedure is to ease the word matching between a patient's request and the system stored information. Notice that NLP does not influence the reasoning engine nor its outcome (i.e., the resulting arguments and their status), it only facilitates the matching operation.

¹⁰ <https://git.ecdf.ed.ac.uk/nkokciya/explainable-argumentation>

¹¹ <https://github.com/FCast07/EQRbot>

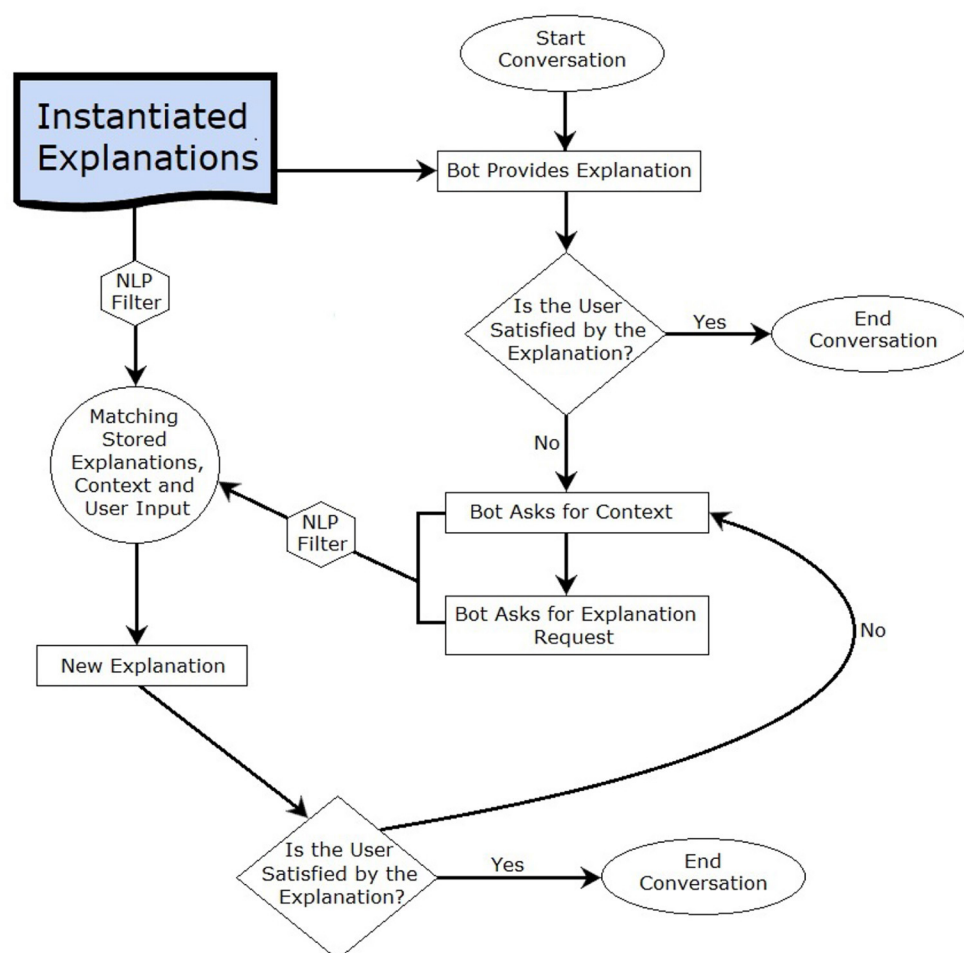


FIGURE 2
Flow chart describing the high-level operations performed by the chatbot (EQRbot).

5.2. The algorithm

The EQRbot's inner operations can be described by an algorithm, [Algorithm 1](#), that takes as input the *Instantiated Explanations* repository (EXP), along with the set of all the possible user queries (Q) related to the data conveyed by the initially provided EQR_explanation (which is also an element of EXP). The procedure continues until the depletion of all the possible queries of Q, that is to say until the user is satisfied with the received information.

Intuitively, `NLP_filter` corresponds to the function that performs a series of Natural Language Process operations as outlined in 5.1. `double_layer_matcher`, instead, represents the BoW similarity procedure in charge of identifying the appropriate response to be delivered. `double_layer_matcher` takes advantage of the context designation, the frequency of key terms occurrence and multiple cross-counts of the input words and the system stored data. Each resulting explanation will then be printed and displayed in the chatbot graphical user interface (GUI).

Proposition 1. *Given the interacting user collaboration (i.e., no out-of-context, non-sense or out-of-the-system-capability input), [Algorithm 1](#) is both sound and complete.*

Indeed, the procedure can provide the requested information that is correct according to the user's input (soundness), and all such answers can be conveyed by the algorithm (completeness). Obviously, this is limited by the data held by the system at the time of the explanation delivery. That is to say, the procedure can only generate explanations determined by the information saved in the system's knowledge base.

Proof.

- [Soundness] The chatbot retrieves the patient's prompt (q) as a pair of context (c) and request (r). Then, the function `find_specific_explanation` (lines 8–23) matches the input with one of the explanations stored in the system (EX) according to a BoW similarity procedure denoted `double_layer_matcher` (lines 16–18). The result of this operation will then consist of the information requested by the user. In case of a mismatch, the process can be repeated until the user's satisfaction (lines 2–5).

```

Input: EXP, an EQR_explanation, and the (finite)
set of the possible user's queries Q
Output: all the requested explanations
1: print(EQR_explanation)
2: for each q ∈ Q:
3:   q == (c, r) ## q is a pair composed
   by a context (c) and specific request (r) ##
4:   find_specific_explanation(q)
5: end for each
6: .
7: .
8: Function find_specific_explanation(q)
9:   NLP_filter(c)
10:  NLP_filter(r)
11:  specific_explanation = " "
12:  similarity_counter = 0
13:  provisional_explanation = " "
14:  for each EX ∈ EXP \ {EQR_explanation}
15:    NLP_filter(EX)
16:    if double_layer_matcher(c, r, EX)
      > similarity_counter then
17:      similarity_counter
      = double_layer_matcher(c, r, EX)
18:      provisional_explanation = EX
19:    endif
20:  end for each
21:  specific_explanation
  = provisional_explanation
22:  print(specific_explanation)
23: end Function

```

Algorithm 1. Matching Queries/Explanations.

- [Completeness] All the requested information can be conveyed by the algorithm. Indeed, each additional explanation the patient might require (associated with the initial EQR explanation) is already saved in the system. They can all be retrieved with the corresponding query (lines 2–5).

□

Since no machine learning operation is involved, hence no time is consumed in training a model, the algorithm will take polynomial time to run. That is because the function `find_specific_explanation` will be called a maximum of $|Q|$ times, i.e., up to the number of elements of Q .

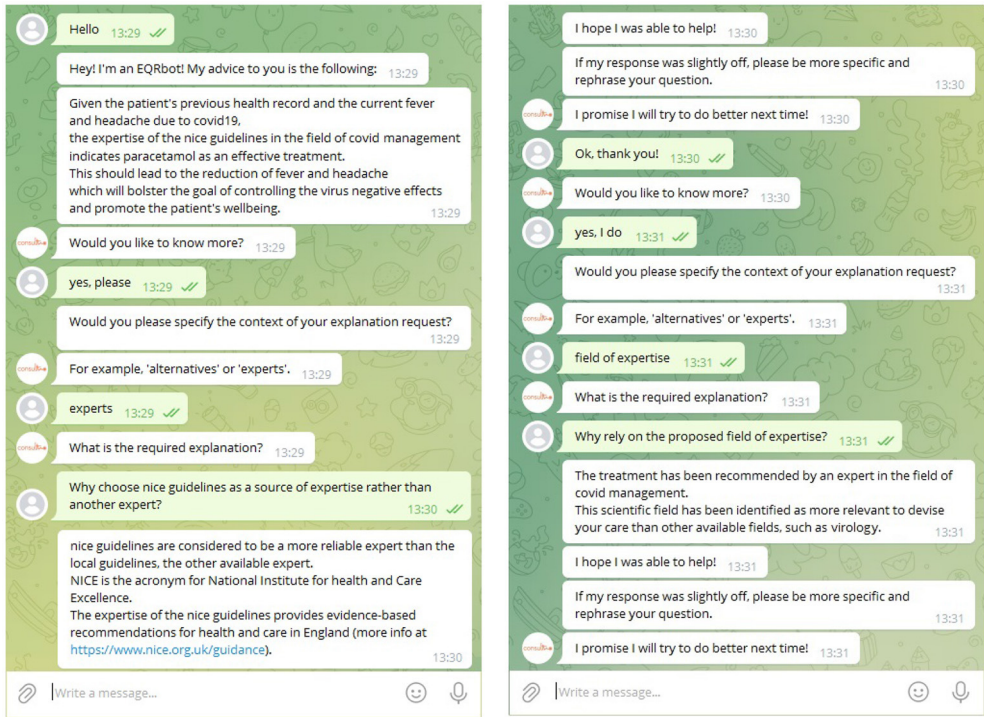
5.3. Implementation

Let us consider the EQR explanation of Example 1. We implemented it *via* a Telegram GUI. We chose to deploy the EQRbot *via* Telegram due to (i) its reputation as one of the most well-known and utilized instant messenger applications, and (ii) its programmer-friendly BOT API. To clarify the interaction depicted in Figure 3, let us suppose that the user monitored by the CONSULT system is, once again, Frida. The electronic health

record supplies the cDSS with two pieces of information: the patient is pregnant, and she is currently suffering from fever and headache caused by the COVID-19 virus. To ease Frida from the pain, when prompted, the CONSULT reasoning engine computes an acceptable (as per Definition 1) piece of advice in the form of an EQR explanation. The EQRbot will display such a recommendation while encouraging also to ask for more details. Supplying the context and the specific request, the patient will demand the rationale behind the choice of the expert that provides the received clinical advice. The chatbot reply involves a natural language explanation based on the acceptable instantiation of the *AS for Expert Reliability* (Figure 3A). In the example, the system considers NICE guidelines as the most reliable source and provides an explanation accordingly. Notice, however, that CONSULT is engineered as a cDSS that supplies recommendations attained from general health guidelines (e.g., NICE). As explicitly stated before its usage, since the system is not supposed to handle conflicts that require professional medical knowledge to be solved, the users should seek advice from their general practitioners would such a circumstance occur. Indeed, this may cause significant harm to the patient if not handled correctly, as emphasized in Snaith et al. (2021). For the same reason, the cDSS (hence the EQRbot) is also updated by the patient's latest wellness sensor readings, the data in their EHR (so, for example, it will not recommend a therapy that has caused negative side effects in the past) and their preferences regarding treatments. The conversation continues in Figure 3B, where Frida interrogates the chatbot for additional information regarding the relevance of the selected field of expertise to the proposed recommendation. Similarly to its previous reply, the bot will formulate an explanation based upon the acceptable instantiation of the *AS for Relevant Field of Expertise*. To completely satisfy the patient's need for clarification, the chatbot will have to output one last explanation, this time about the acceptable instantiation of the *AS for alternative Clinical Options*. Indeed, the patient desires to know if alternative treatments are available (because, for instance, the drug indicated by CONSULT is not currently accessible to her). However, the cDSS confirms its previous recommendation informing Frida that, due to her pregnancy, paracetamol is the most appropriate remedy to assume (Figure 3C).

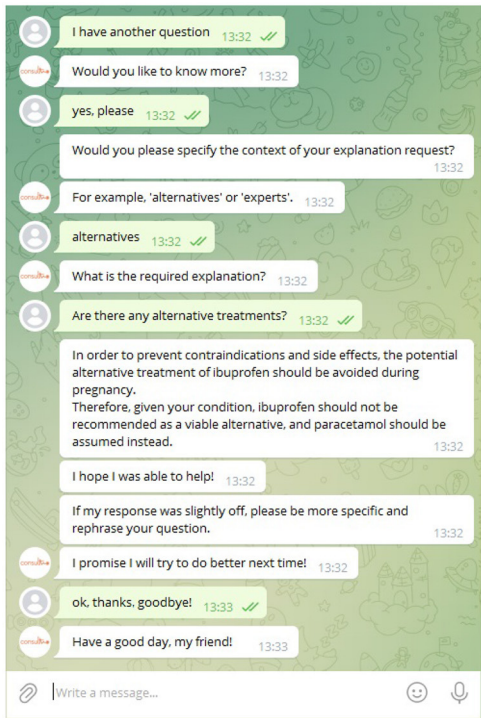
5.4. Evaluating the EQRbot against the CONSULT baseline

A seven day within-subjects mixed-methods run in-the-wild (Waterson et al., 2002) study has been conducted to assess the usability and acceptability of the CONSULT system with two different versions: with and without a chatbot. Such a pilot study demonstrated that real users could employ the application over an extended period (Balatsoukas et al., 2020). Connie, the conversational agent previously equipped with the cDSS at the time of the experiment, accommodates the patients willing to seek immediate evidence-based advice about a specific health problem. Informed by the user's vital data, preferences, EHR and clinical guidelines retrieved by the CONSULT system, the chatbot provides any additional explanation regarding the proposed



A ASEX example

B ASF example



C ASCO example

FIGURE 3
Instance of a conversation with the EQRbot starting from the explanation of Example 1. The displayed interaction captures the patient inquiries regarding the involved expert (A), the field of expertise (B) and possible alternatives to the proposed treatment (C). Matching the user's input, context and the information stored in the system, the EQRbot provides the additional requested explanation via the acceptable instantiations of the, respectively, ASEX, ASF, and ASCO schemes.

recommendation. The main aspects that characterize Connie can be outlined as:

- *User's Input.* No free interaction occurs since the user's prompt is restricted to hard-coded multiple options.
- *Interface.* The chat, and related conversation log, are graphically displayed *via* Mattermost¹².
- *Chatbot Type.* Connie is a rule-based chatbot¹³, i.e., an agent capable of responding only by following predetermined (scripted) replies according to the user's input.
- *Reasoning Engine.* The bot leverages the results of the operations performed by the CONSULT system by means of the computational argumentation solver ASPARTIX.
- *Explanation Delivery.* No particular strategy is deployed. The explanations are triggered *via* the options selected by the user.

An example of a conversation with Connie is illustrated in Figure 4B. Here the interacting patient is given the choice of selecting among four different options in response to the question "What can I help you with?". The user then decides to report a symptom concerning backpain, asking also for more details once a reply is given. This option triggers one last response from the chatbot, thus providing the explanation behind the rationale of the proposed recommendation. Nonetheless, Connie presents some limitations, as summarized by the result of the pilot study: "[...] the lack of a more natural conversation flow when interacting with the chatbot (e.g., close to the one that they [the patients] would have with their GP)" (Balatsoukas et al., 2020).

Against Connie, considered as the previous baseline, the EQRbot yields several advantages, as highlighted by the comparative table of Figure 4A:

- *User's Input.* Free textual interaction. Each user's prompt will be parsed by the chatbot NLP filter and matched with the most appropriate reply. Any non-sense or out-of-context input will be addressed by a random response from the bot.
- *Interface.* The chat, and related conversation log, are graphically displayed *via* Telegram.¹⁴
- *Chatbot Type.* EQRbot is a retrieval-based chatbot, i.e., an agent that mostly retrieves its replies from a database of potential responses according to the most relevant match with the user's input.
- *Reasoning Engine.* The bot leverages the results of the operations performed by the CONSULT system by means of the computational argumentation solver ASPARTIX.
- *Explanation Delivery.* The aim is to reduce the number of potential user queries (including possible follow-on questions) and concerns by concentrating the most relevant information about a specific recommendation within a single explanation, i.e., the one elicited by an acceptable instantiation of the EQR scheme.

The EQRbot represent an improvement over Connie since it addresses (in four out of the five listed main features) the shortcomings ensuing from the pilot study outcome. Indeed, it allows for (i) better approximations of natural conversations without textual restriction, by employing (ii) Telegram GUI, i.e., a more user-friendly, and popular messaging application than Mattermost. In general, (iii) retrieval-based chatbots are more versatile and flexible than rule-based ones, hence more suited for real-world exchange of arguments. Finally, despite its simplicity, (iv) having an explanation strategy bring the EQRbot closer to an authentic question-answer dialog.

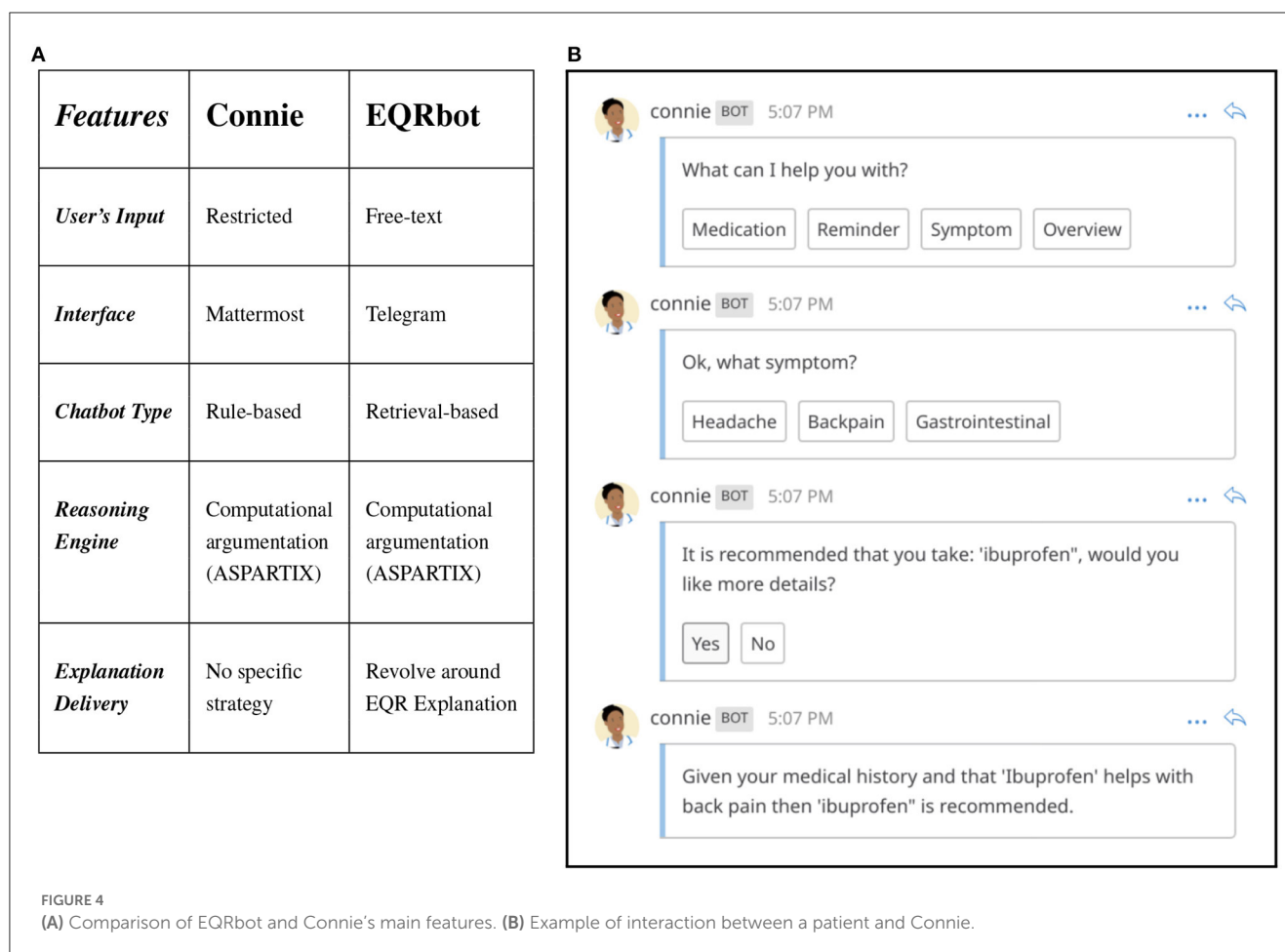
6. Discussion

Although argument schemes-based clinical explanations have already been employed in studies such as Atkinson et al. (2006), Kökciyan et al. (2020), Shaheen et al. (2021), and Sassoon et al. (2021), the EQR scheme proposed herein emerges as a model designed to efficiently deliver a significant amount of information (both practical and theoretical) at once. Indeed, EQR explanations constitute the core notions around which all the data, possibly required by subsequent follow-on queries, are clustered into user-friendly natural language snippets of texts. Nevertheless, the envisaged implementation (of which Figure 3 represents a very restricted example) of this new argument scheme *via* the EQRbot presents some limitations, the most prominent of which concerns the delivery of the explanations. The conversation that occurs with the patient, albeit simple and clear, lacks a fully-fledged formal protocol with a complete set of available locutions, tracked utterances commitment store, precise semantics and pragmatics (McBurney and Parsons, 2009). A protocol for an explanation dialog has been given in Bex and Walton (2016) with a complete list of locutions. However, to evaluate the provided explanation, the explainee needs to resort to a different dialog protocol (denoted examination). Similarly, Madumal et al. (2019) devise a study for modeling explanation dialogs by following a data-driven approach. The resulting formalization embeds (possibly several) argumentation dialogs nested in the outer layer of the explanation protocol. Finally, also the dialog structure proposed (for a previous version of the CONSULT chatbot) by Sassoon et al. (2019) in the context of explanations for wellness consultation exploits multiple dialog types (e.g., *persuasion*, *deliberation* and *information seeking*) and their respective protocols whilst mostly focusing on the course of action to undertake. This is different from the anticipated EQR dialog (sketched in McBurney and Parsons, 2021 as Explanation-Question-Response), whose protocol is halfway between *persuasion*, *information-giving/seeking* and *query* and more comprehensively incorporates locutions for handling each of these tasks without the need for adopting a *control layer* (Cogan et al., 2006) or switching between protocols. This allows for a simpler formalization and more genuine dialogs. For all of these reasons, future implementations of EQRbot will provide for the addition of a formal protocol and an adjustment to the chatbot's memory. That is to say, the bot's capability for recalling the arguments previously moved in the conversation and recorded in the commitment store. Indeed, considering that the EQR explanations have been informed by several CQs that

¹² <https://mattermost.com/>

¹³ <https://www.codecademy.com/article/what-are-chatbots>

¹⁴ <https://telegram.org/>



should comprehend all the possible challenges moved to them, no problem will arise if the user's inquiries regard these explanations or their specifics. However, if the inquiries concern a reference to an argument that occurred in an earlier stage of the dialog, the chatbot may not be able to properly address the request.

The landscape of argumentation-based chatbots has seen an increase in interest in recent years. For example, ArguBot (Bistarelli et al., 2021), developed using Google DialogFlow, employs ASPARTIX to compute arguments from an underlying Bipolar AF, or BAF, (Cayrol and Lagasquie-Schiex, 2005) to support or challenge the user's opinion about a dialog topic. The conversational capabilities of ArguBot are, however, restricted by the arguments stored in the BAF as its knowledge base, limiting its dialectical potential only to specific fully-developed interactions. One of the main problems concerning argumentation-based chatbots is indeed the creation of a proper knowledge base from which the bot's arguments can be retrieved and employed to interact with the user. The research of Chalaguine et al. (2018) and Chalaguine and Hunter (2018, 2019) outline *harvesting* and *crowd-sourcing* methodologies capable of collecting arguments and counter-arguments on a specific topic, thus generating suitable and persuasive knowledge bases for chatbots [e.g., Chalaguine and Hunter (2020)], and, harnessing also hand-crafted counterarguments due to the topic sensitivity, Chalaguine

and Hunter (2021)]. Unlike the studies presented thus far, the knowledge base of the EQRbot is personalized on the patient's preferences and health data. That information is constantly updated, making it possible to generate a potentially indefinite number of diverse explanations (although the user will need to restart the conversation to allow for the acquisition of the modified knowledge base, since the EQRbot cannot alter its stored responses during an interaction). Finally, although still resorting to similarity algorithms to retrieve appropriate arguments from a fixed knowledge base, Fazzinga et al. (2021) designed a bot that performs a reasoning step with multiple elements of user information before outputting each reply. Notice, however, that our EQRbot already performs such a step before selecting the final answer. Indeed, the list of responses fed to the chatbot is the result of a computation of the framework's acceptable arguments generated from the data and templates presented in the CONSULT system. Restarting the conversation with the EQRbot before each new explanation request will ensure that a new reasoning process (that involves the overall AF) will take place.

Lastly, further improvements could also arise by combining the recent developments in the field of *Argument Mining* (Cabrio and Villata, 2018) with additional chatbot code-based instructions. The swift generation of AFs comprising domain-specific arguments can indeed assist the bot in performing engaging dialogs such

that the user's claims might be constructively challenged by more persuasive and precise explanations. The mining should occur from a specialized dataset composed of annotated clinical abstracts as in Mayer et al. (2020) or Stylianou and Vlahavas (2021), where the authors provide a complete argument mining pipeline capable of classifying argument components as *evidence/claim* and argument relations as *attack/support*. In addition, the research presented in Mayer et al. (2021) extends the pipeline by detecting also the effects on the outcome associated with the identified argumentative components.

6.1. Planned user study

To fully evaluate the EQRbot performances, we are currently planning a user study. The goal of the study is to analyze the interactions between the patients and the chatbot, such as how often a conversation is initiated, how long the question/answer session is on average and which are the most common queries prompted by the user. In particular, we are interested in a qualitative assessment of the provided explanations and the general level of users' satisfaction toward them. As discussed before, CONSULT handles data from patients' Electronic Health Records and suggests treatments (following clinical guidelines and stakeholders' preferences) that have already been tested on the interacting subjects, thus preventing any contraindications or side effects. Therefore the recommendations and potential explanations delivered by the EQRbot will not risk harming the user, and will instead indicate to contact medical professionals when required. However, if such a message occurs frequently, this may have the negative consequence of raising distrust from the patient against the system which may then overlook such a recommendation hence precluding (possibly essential) communications with the main caregivers. For this reason, the participants of the study will be preemptively informed of the cDSS limitations and its main functions. In addition, they will also receive a user manual to be examined whenever needed. The study is expected to last for two weeks, during which the patients are free to explore the system functionalities and interact with the chatbot. Before the beginning of the experiment, the participants will be interviewed in order to understand what they seek and prospect from the interactions with the cDSS and the EQRbot. A similar interview will also be conducted at the end of the study, where it will be possible to compare the user experience with their initial expectations and where feedback for further improvements will be collected.

7. Conclusion

Designed as a model capable of efficiently delivering both practical and theoretical information during inter-agent (human or AI) explanations, the EQR argument scheme proposed herein formalizes the consequences yielded (and the presumptive reasoning leading to them) by acting upon an expert opinion. In this paper, we outlined an approach that integrates the EQR scheme in the current research landscape involving decision support systems and argument-based explanations. In particular, we have focussed on studies regarding medical applications of

such reasoning patterns, and we have presented a possible way of enhancing the related explanation templates. Indeed, one of the main advantages offered by the provided contributions is the incorporation of clinically specialized AS (e.g., ASPT) into the newly detailed EQR scheme structure. This will give more opportunities for inquiry to an agent seeking clarification since there are more aspects that can be interrogated and that can help in finding a satisfactory and more complete explanation. For example, which expert is informing the suggested treatment is a piece of information that might increase the patients' trust in the medical recommendation system. Furthermore, we have presented an implementation of the proposed contributions by equipping the CONSULT cDSS with a chatbot that employs acceptable EQR scheme instantiations as the core element to convey explanations. This is a substantial contribution to the research field of argumentation-based human-agent interactions. Indeed, our bot is guided exclusively by an argumentation reasoning engine in its decision-making process while it converses with the user: no machine learning algorithm is involved in the procedure. In addition, NLP is utilized only as a means for enhancing the word matching between the user input (which is completely free and not limited to multiple choice options) and the system stored explanations. Unlike other chatbots in the literature, the EQRbot depends upon a dynamic knowledge base that is constantly updated by the patient's data received from the health sensors and their EHR. This entails more personalized and, possibly, disparate interactions, as long as the user restarts the conversation (which will allow the reasoning engine to generate new explanations upon the updated knowledge base). Finally, we deploy our bot *via* Telegram. Such a choice ensures a convenient programmer API along with a well-known and user-friendly GUI.

Data availability statement

The provided link: <https://github.com/FCast07/EQRbot> refers to the GitHub repository that stores the chatbot programming code.

Author contributions

FC: main idea, first draft, and chatbot implementation. AG: telegram GUI for the chatbot. PM, SP, and IS: conceptualization, edit, and review. ES: edit and review. All authors contributed to the article and approved the submitted version.

Funding

This research was partially funded by the UK Engineering & Physical Sciences Research Council (EPSRC) under Grant #EP/P010105/1.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

References

- Antaki, C., and Leudar, I. (1992). Explaining in conversation: towards an argument model. *Eur. J. Soc. Psychol.* 22, 181–194.
- Atkinson, K., and Bench-Capon, T. (2007). Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artif. Intell.* 171, 855–874. doi: 10.1016/j.artint.2007.04.009
- Atkinson, K., Bench-Capon, T., and Modgil, S. (2006). "Argumentation for decision support," in *International Conference on Database and Expert Systems Applications*, eds S. Bressan, J. Küng, and R. Wagner (Berlin; Heidelberg: Springer), 822–831. doi: 10.1007/11827405_80
- Balatsoukas, P., Porat, T., Sassoon, I., Essers, K., Kokciyan, N., Chapman, M., et al. (2019). "User involvement in the design of a data-driven self-management decision support tool for stroke survivors," in *IEEE EUROCON 2019-18th International Conference on Smart Technologies* (Novi Sad: IEEE), 1–6. doi: 10.1109/EUROCON.2019.8861812
- Balatsoukas, P., Sassoon, I., Chapman, M., Kokciyan, N., Drake, A., Modgil, S., et al. (2020). "In the wild pilot usability assessment of a connected health system for stroke self management," in *2020 IEEE International Conference on Healthcare Informatics (ICHI)* (Oldenburg: IEEE), 1–3. doi: 10.1109/ICHI48887.2020.9374338
- Bellotti, V., and Edwards, K. (2001). Intelligibility and accountability: human considerations in context-aware systems. *Hum. Comput. Interact.* 16, 193–212. doi: 10.1207/S15327051HCI16234_05
- Bench-Capon, T. J., and Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artif. Intell.* 171, 619–641. doi: 10.1016/j.artint.2007.05.001
- Besnard, P., and Hunter, A. (2008). *Elements of Argumentation*, Vol. 47. Cambridge: MIT Press.
- Bex, F., and Walton, D. (2016). Combining explanation and argumentation in dialogue. *Argument Comput.* 7, 55–68. doi: 10.3233/AAC-160001
- Bilici, E., Despotou, G., and Arvanitis, T. N. (2018). The use of computer-interpretable clinical guidelines to manage care complexities of patients with multimorbid conditions: a review. *Digital Health* 4, 2055207618804927. doi: 10.1177/2055207618804927
- Bistarelli, S., Taticchi, C., and Santini, F. (2021). "A Chatbot Extended with Argumentation," in *Proceedings of the 5th Workshop on Advances in Argumentation in Artificial Intelligence 2021 co-located with the 20th International Conference of the Italian Association for Artificial Intelligence (AIXIA 2021)*, eds M. D'Agostino, F. A. D'Asaro, and C. Larese (Milan: CEUR Workshop Proceedings).
- Bondarenko, A., Dung, P. M., Kowalski, R. A., and Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. *Artif. Intell.* 93, 63–101.
- Cabrio, E., and Villata, S. (2018). "Five years of argument mining: a data-driven analysis," in *IJCAI, Vol. 18* (Stochholm), 5427–5433.
- Castagna, F., Parsons, S., Sassoon, I., and Sklar, E. I. (2022). "Providing explanations via the EQR argument scheme," in *Computational Models of Argument: Proceedings of COMMA 2022*. (Cardif: IOS Press), 351–352.
- Cayrol, C., and Lagasque-Schiex, M.-C. (2005). "On the acceptability of arguments in bipolar argumentation frameworks," in *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty ECSQARU 2005*, ed L. Godo (Berlin; Heidelberg: Springer), 378–389. doi: 10.1007/11518655_33
- Chalaguine, L., Hamilton, F., Hunter, A., and Potts, H. (2018). "Argument harvesting using chatbots," in *Proceedings of COMMA* (Warsaw), 149.
- Chalaguine, L., and Hunter, A. (2018). "Chatbot design for argument harvesting," in *Computational Models of Argument: Proceedings of COMMA 2018*, eds S. Modgil, K. Budzynska, and J. Lawrence (Warsaw: IOS Press), 457.
- Chalaguine, L., and Hunter, A. (2021). "Addressing popular concerns regarding COVID-19 vaccination with natural language argumentation dialogues," in *European Conference on Symbolic and Quantitative Approaches with Uncertainty, ECSQARU 2021* (Prague: Springer-Verlag), 59–73. doi: 10.1007/978-3-030-86772-0_5
- Chalaguine, L. A., and Hunter, A. (2019). "Knowledge acquisition and corpus for argumentation-based chatbots," in *Proceedings of the 3rd Workshop on Advances In Argumentation In Artificial Intelligence co-located with the 18th International Conference of the Italian Association for Artificial Intelligence* (Rende: CEUR Workshop Proceedings), 1–14.
- Chalaguine, L. A., and Hunter, A. (2020). "A persuasive chatbot using a crowd-sourced argument graph and concerns," in *Computational Models of Argument: Proceedings of COMMA 2020* (Perugia), 9.
- Chapman, M., Abigail, G., Sassoon, I., Kökciyan, N., Sklar, E. I., Curcin, V., et al. (2022). "Using microservices to design patient-facing research software," in *2022 IEEE 18th International Conference on e-Science (e-Science)* (Salt Lake City, UT: IEEE), 44–54. doi: 10.1109/eScience55777.2022.00019
- Chapman, M., Balatsoukas, P., Kökciyan, N., Essers, K., Sassoon, I., Ashworth, M., et al. (2019). "Computational argumentation-based clinical decision support," in *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems. Vol. 4* (Montreal, QC: International Foundation for Autonomous Agents and Multiagent Systems), 2345–2347.
- Chesnevar, C. I., Maguitman, A. G., and Loui, R. P. (2000). Logical models of argument. *ACM Comput. Surv.* 32, 337–383. doi: 10.1145/371578.371581
- Cogan, E., Parsons, S., and McBurney, P. (2006). "New types of inter-agent dialogues," in *Argumentation in Multi-Agent Systems*, eds S. Parsons, N. Maudet, P. Moraitis, and I. Rahwan (Berlin; Heidelberg: Springer), 154–168. doi: 10.1007/11794578_10
- Cyras, K., Delaney, B., Prociuk, D., Toni, F., Chapman, M., Domínguez, J., et al. (2018). "Argumentation for explainable reasoning with conflicting medical recommendations," in *CEUR Workshop Proceedings* (Tempe, FL).
- Cyras, K., Rago, A., Albini, E., Baroni, P., and Toni, F. (2021). "Argumentative XAI: a survey," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence* (Montreal).
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77, 321–357.
- Egly, U., Gaggli, S. A., and Woltran, S. (2008). "Aspartix: implementing argumentation frameworks using answer-set programming," in *Logic Programming*, eds M. Garcia de la Band and E. Pontelli (Berlin; Heidelberg: Springer), 734–738. doi: 10.1007/978-3-540-89982-2_67
- Essers, K., Chapman, M., Kokciyan, N., Sassoon, I., Porat, T., Balatsoukas, P., et al. (2018). "The CONSULT system: Demonstration," in *Proceedings of the 6th International Conference on Human-Agent Interaction (HAI '18)* (New York, NY: Association for Computing Machinery), 385–386. doi: 10.1145/3284432.3287170
- Fairweather, E., Wittner, R., Chapman, M., Holub, P., and Curcin, V. (2020). "Non-repudiable provenance for clinical decision support systems," in *Provenance and Annotation of Data and Processes: 8th and 9th International Provenance and Annotation Workshop, IPAW 2020 + IPAW 2021, Virtual Event, July 19–22, 2021, Proceedings* (Springer-Verlag), 165–182. doi: 10.1007/978-3-030-80960-7_10
- Fazzinga, B., Galassi, A., and Torroni, P. (2021). "An argumentative dialogue system for COVID-19 vaccine information," in *ILogic and Argumentation: 4th International Conference, CLAR 2021, Hangzhou, China, October 20–22, 2021, Proceedings* (Springer-Verlag), 477–485.
- Grando, M. A., Moss, L., Sleeman, D., and Kinsella, J. (2013). Argumentation-logic for creating and explaining medical hypotheses. *Artif. Intell. Med.* 58, 1–13. doi: 10.1016/j.artmed.2013.02.003
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). XAI-explainable artificial intelligence. *Sci. Robot.* 4, eaay7120. doi: 10.1126/scirobotics.aay7120
- Kökciyan, N., Chapman, M., Balatsoukas, P., Sassoon, I., Essers, K., Ashworth, M., et al. (2019). A collaborative decision support tool for managing chronic conditions. *Stud. Health. Technol. Inform.* 264, 644–648. doi: 10.3233/SHTI190302
- Kökciyan, N., Parsons, S., Sassoon, I., Sklar, E., and Modgil, S. (2020). "An argumentation-based approach to generate domain-specific explanations," in *Multi-Agent Systems and Agreement Technologies, 17th European Conference on Multi-Agent Systems, EUMAS 2020, and 7th International Conference on Agreement Technologies, AT 2020* (Thessaloniki: Springer), 319–337.
- Kökciyan, N., Sassoon, I., Sklar, E., Modgil, S., and Parsons, S. (2021). Applying metalevel argumentation frameworks to support medical decision making. *IEEE Intell. Syst.* 36, 64–71. doi: 10.1109/MIS.2021.3051420

- Kokciyan, N., Sassoan, I., Young, A. P., Chapman, M., Porat, T., Ashworth, M., et al. (2018). "Towards an argumentation system for supporting patients in self-managing their chronic conditions," in *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, Health Intelligence Workshop at AAAI Conference on Artificial Intelligence* (New Orleans, LO: AAAI Press), 455–462.
- Lindgren, H., Kampik, T., Rosero, E. G., Blusi, M., and Nieves, J. C. (2020). Argumentation-based health information systems: a design methodology. *IEEE Intell. Syst.* 36, 702–780. doi: 10.1109/MIS.2020.3044944
- Madumal, P., Miller, T., Sonenberg, L., and Vetere, F. (2019). A grounded interaction protocol for explainable artificial intelligence. *arXiv [Preprint]*. arXiv: 1903.02409. doi: 10.48550/arXiv.1903.02409
- Mayer, T., Cabrio, E., and Villata, S. (2020). "Transformer-based argument mining for healthcare applications," in *ECAI 2020* (Santiago de Compostela: IOS Press), 2108–2115.
- Mayer, T., Marro, S., Cabrio, E., and Villata, S. (2021). Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials. *Artif. Intell. Med.* 118, 102098. doi: 10.1016/j.artmed.2021.102098
- McBurney, P., and Parsons, S. (2009). "Dialogue games for agent argumentation," in *Argumentation in Artificial Intelligence*, eds G. Simari and I. Rahwan (Boston, MA: Springer), 261–280. doi: 10.1007/978-0-387-98197-0
- McBurney, P., and Parsons, S. (2021). Argument schemes and dialogue protocols: Doug Walton's legacy in artificial intelligence. *J. Appl. Log.* 8, 263–286.
- Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007
- Modgil, S., and Prakken, H. (2013). A general account of argumentation with preferences. *Artif. Intell.* 195, 361–397. doi: 10.1016/j.artint.2012.10.008
- Oliveira, T., Dauphin, J., Satoh, K., Tsumoto, S., and Novais, P. (2018). "Argumentation with goals for clinical decision support in multimorbidity," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm).
- Pollock, J. L. (1987). Defeasible reasoning. *Cogn. Sci.* 11, 481–518.
- Reiter, R. (1980). A logic for default reasoning. *Artif. Intell.* 13, 81–132.
- Sassoan, I., Kökciyan, N., Modgil, S., and Parsons, S. (2021). Argumentation schemes for clinical decision support. *Argument Comput.* 12, 329–355. doi: 10.3233/AAC-200550
- Sassoan, I., Kökciyan, N., Sklar, E., and Parsons, S. (2019). "Explainable argumentation for wellness consultation," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: First International Workshop, EXTRAAMAS 2019, Montreal, QC, Canada, May 13–14, 2019* (Berlin; Heidelberg: Springer), 186–202. doi: 10.1007/978-3-030-30391-4_11
- Shaheen, Q., Toniolo, A., and Bowles, K. F. (2021). "Argumentation-based explanations of multimorbidity treatment plans," in *PRIMA 2020: Principles and Practice of Multi-Agent Systems: 23rd International Conference, Nagoya, Japan, November 18–20, 2020, Proceedings. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, eds T. Uchiya, Q. Bai, and I. M. Maestre (Cham: Springer), 394–402. doi: 10.1007/978-3-030-69322-0_29
- Snaith, M., Nielsen, R. Ø., Kotnis, S. R., and Pease, A. (2021). Ethical challenges in argumentation and dialogue in a healthcare context. *Argument Comput.* 12, 249–264. doi: 10.3233/AAC-200908
- Stylianou, N., and Vlahavas, I. (2021). Transformed: end-to-end transformers for evidence-based medicine and argument mining in medical literature. *J. Biomed. Inform.* 117, 103767. doi: 10.1016/j.jbi.2021.103767
- Sutton, R. T., Pincok, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., and Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit. Med.* 3, 1–10. doi: 10.1038/s41746-020-0221-y
- Tolchinsky, P., Modgil, S., Atkinson, K., McBurney, P., and Cortés, U. (2012). Deliberation dialogues for reasoning about safety critical actions. *Auton. Agents Multi Agent Syst.* 25, 209–259. doi: 10.1007/s10458-011-9174-5
- Toni, F. (2014). A tutorial on assumption-based argumentation. *Argument Comput.* 5, 89–117. doi: 10.1080/19462166.2013.869878
- Vasey, B., Ursprung, S., Beddoe, B., Taylor, E. H., Marlow, N., Bilbro, N., et al. (2021). Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. *JAMA Netw. Open* 4, e211276. doi: 10.1001/jamanetworkopen.2021.1276
- Vassiliades, A., Bassiliades, N., and Patkos, T. (2021). Argumentation and explainable artificial intelligence: a survey. *Knowledge Eng. Rev.* 36, e5. doi: 10.1017/S0269888921000011
- Vilone, G., and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inform. Fusion* 76, 89–106. doi: 10.1016/j.inffus.2021.05.009
- Walton, D. (1996). *Argumentation Schemes for Presumptive Reasoning*. New York, NY: Routledge.
- Walton, D. (1997). *Appeal to Expert Opinion: Arguments from Authority*. Pennsylvania State University Press, University Park, PA.
- Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511802034
- Waterson, S., Landay, J. A., and Matthews, T. (2002). "In the lab and out in the wild: remote web usability testing for mobile devices," in *CHI'02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)* (New York, NY: Association for Computing Machinery), 796–797. doi: 10.1145/506443.506602



OPEN ACCESS

EDITED BY

Emmanuelle Dietz,
Airbus, Germany

REVIEWED BY

Nick Bassiliades,
Aristotle University of Thessaloniki, Greece
Alfonso Guarino,
University of Foggia, Italy

*CORRESPONDENCE

Antonio Rago
✉ a.rago@imperial.ac.uk

SPECIALTY SECTION

This article was submitted to
Machine Learning and Artificial Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

RECEIVED 15 November 2022

ACCEPTED 20 March 2023

PUBLISHED 06 April 2023

CITATION

Albini E, Rago A, Baroni P and Toni F (2023)
Achieving descriptive accuracy in explanations
via argumentation: The case of probabilistic
classifiers. *Front. Artif. Intell.* 6:1099407.
doi: 10.3389/frai.2023.1099407

COPYRIGHT

© 2023 Albini, Rago, Baroni and Toni. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Achieving descriptive accuracy in explanations *via* argumentation: The case of probabilistic classifiers

Emanuele Albini¹, Antonio Rago^{1*}, Pietro Baroni² and
Francesca Toni¹

¹Department of Computing, Imperial College London, London, United Kingdom, ²Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Brescia, Italy

The pursuit of trust in and fairness of AI systems in order to enable human-centric goals has been gathering pace of late, often supported by the use of *explanations* for the outputs of these systems. Several properties of explanations have been highlighted as critical for achieving trustworthy and fair AI systems, but one that has thus far been overlooked is that of *descriptive accuracy* (DA), i.e., that the explanation contents are in correspondence with the internal working of the explained system. Indeed, the violation of this core property would lead to the paradoxical situation of systems producing explanations which are not suitably related to how the system actually works: clearly this may hinder user trust. Further, if explanations violate DA then they can be deceitful, resulting in an unfair behavior toward the users. Crucial as the DA property appears to be, it has been somehow overlooked in the XAI literature to date. To address this problem, we consider the questions of formalizing DA and of analyzing its satisfaction by explanation methods. We provide formal definitions of *naive*, *structural* and *dialectical* DA, using the family of probabilistic classifiers as the context for our analysis. We evaluate the satisfaction of our given notions of DA by several explanation methods, amounting to two popular feature-attribution methods from the literature, variants thereof and a novel form of explanation that we propose. We conduct experiments with a varied selection of concrete probabilistic classifiers and highlight the importance, with a user study, of our most demanding notion of dialectical DA, which our novel method satisfies by design and others may violate. We thus demonstrate how DA could be a critical component in achieving trustworthy and fair systems, in line with the principles of human-centric AI.

KEYWORDS

argumentation, descriptive accuracy, explainable AI, probabilistic classifiers, properties

1. Introduction

Equipping automated decision systems with explanation capabilities is a compelling need which lies at the basis of the rapid growth of the research field of explainable AI (XAI) in recent years (Guidotti et al., 2019) and is receiving an increasing attention from government and regulatory bodies, like the European Commission. Quoting the report of the [Independent high-level expert group on Artificial Intelligence set up by the European Commission](#) (2019): “Whenever an AI system has a significant impact on people’s lives, it should be possible to demand a suitable explanation of the AI system’s decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g., layperson, regulator or researcher).”

By providing explanations, a system goes beyond just presenting its outcomes as oracles: rather, they are subjected to the scrutiny of the cognitive capabilities of the users, who receive means to achieve a better understanding of the reasons underlying system's decisions and/or of its internal operation. In this way, the adoption of an active and conscious role by users is supported: they will be able to criticize or agree with system decisions, based on a cognitively elaborated motivation, rather than blindly rejecting or accepting them. Thus, explanations for the outputs of AI systems are widely understood as crucial to support trust in these systems (Ribeiro et al., 2016; Miller, 2019; Zerilli et al., 2022).

Due to their role in promoting users' understanding and involvement, it is no surprise that the two properties of *cognitive explainability and transparency* are widely regarded as key factors and technical challenges of Human-Centric AI, as evidenced in the introductory article of this special issue (Kakas et al., 2022). For instance, "Make clear why the system did what it did" is one of the design guidelines for human-AI interaction presented by Amershi et al. (2019), while the Research Roadmap of the European network of Human-Centered Artificial Intelligence (www.humane-ai.eu) regards the fact that AI systems are explainable and accountable as a basic prerequisite for human-in-the-loop activities.

This paper contributes to the development of explainability for human-centric AI by proposing a formal treatment of the notion of *descriptive accuracy* (DA), a crucial property for explanations supporting fair AI systems deserving trust, and by showing how DA requirements can be achieved in practice through a suitable form of explanation, called *DARXs* (acronym for *Dialectically Accurate Relational Explanations*). Both the formal treatment of DA and the definition of DARX are based on ideas and formalisms from the field of *Argumentation Theory*, connecting the present contribution to the subject of the special issue. Argumentation theory (also referred to in the literature as *computational* or *artificial* argumentation, e.g., see Atkinson et al., 2017; Baroni et al., 2018 for overviews) has recently been advocated, in a variety of ways, as a mechanism for supporting explainable AI (see Cyras et al., 2021; Vassiliades et al., 2021 for recent surveys). A popular use thereof is as a means for representing the information in an existing AI system in a way which is more amenable for human consumption than typical explanation methods, e.g., as in Timmer et al. (2015) and Rago et al. (2021). This use of argumentation is the inspiration also for this paper: the formulations of DA we propose are defined for abstract notions of explanation inspired by the argumentation frameworks in the seminal works of Toulmin (1958) and Cayrol and Lagasquie-Schiex (2005), DARXs are inspired by bipolar argumentation frameworks (Cayrol and Lagasquie-Schiex, 2005); and our definitions of DA bear resemblance to properties originally proposed (for various forms of argumentation frameworks) by Amgoud and Ben-Naim (2018) and by Baroni et al. (2019).

The paper is organized as follows. Section 2 presents the motivations and contribution of the work, in particular positioning our contribution in the context of the special issue, while Section 3 discusses related works. Then, after providing the required preliminary notions in Section 4, we introduce the proposed formal treatment of DA in Section 5. Section 6 examines the satisfaction of DA by some existing and novel explanation approaches, showing

that, differently from other proposals, DARX guarantees a full satisfaction of DA requirements. These formal results are backed by an experimental evaluation in Section 7 and by a human experiment in Section 8, before concluding in Section 9.

2. Motivations and contribution

Being immersed in the human-centric perspective, the issue of realizing explainable and transparent system does not only represent a challenging and fascinating socio-technical problem to tackle (Miller, 2019), but also involves substantial ethical aspects and requires the satisfaction of human-centric properties, like trustworthiness and fairness.

First, the explanations provided for the outputs of a system are a key factor in achieving user *trust*, a prerequisite for acceptance of the decisions of a system when deemed to be trustworthy. However, as pointed out by Jacovi et al. (2021), trust, which is an attitude of the trustors (in our case, the systems' users), is distinguished from trustworthiness, which is a property of the trustees (in our case the explained systems), i.e., the capability of maintaining some *contract* with the users. In fact, "trust and trustworthiness are entirely disentangled: ... trust can exist in a model that is not trustworthy, and a trustworthy model does not necessarily gain trust" (Jacovi et al., 2021). This makes the goal of achieving trust, and the role of explanations therefor, a rather tricky issue. On the one hand, there can be situations where trust is achieved by explanations which are convincing but somehow deceptive. On the other hand, there can be situations where an otherwise trustworthy system loses users' trust due to problems in its explanations' capabilities, e.g., as pointed out by Jacovi et al. (2021). For illustration, consider the case of an AI-based medical system predicting, for a patient, a high risk of getting disease X and including in the explanation the fact that some parameter Y in the patient's blood test is high. If the patient deems the system trustworthy, they may try to change (if possible) the value of Y, e.g., by lifestyle changes. If they find out that the value of Y was actually irrelevant, i.e., the diagnosis would have been the same with a low value of Y, and thus trying to modify it will not have the intended impact on the system prediction, then the patient's trust will be negatively affected, independently of the correctness of the diagnosis.

Thus, trust in an otherwise accurate system can be hindered or even destroyed by some drawbacks of the explanations it provides.

Trust is however not the only issue at stake. Continuing the illustration, suppose the patient never gets to know about the irrelevance of parameter Y. Then, their trust may be preserved, but then a possibly deceitful system would remain in place. This shows that, in connection to their impact on trust, explanations also have an important role toward *fairness* of AI systems: the description of the principle of *fairness* in the report by the *Independent high-level expert group on Artificial Intelligence set up by the European Commission* (2019) states that "the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice." This indication complements the requirement of "ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatization." Two complementary facets of fairness emerge here. The latter concerns a possible unjust

treatment caused by system biases toward specific user features, while the former addresses the risk that the system may treat its users improperly due to inappropriate design choices for the explanations. This form of unfairness applies to *all* users, rather than just *some*, as in the case of selective system biases.

Avoiding *selectively unfair* (biased) systems is receiving a great deal of attention in the literature (see, for instance, Dwork et al., 2012; Heidari et al., 2019; Hutchinson and Mitchell, 2019; Binns, 2020; Rüz, 2021), whereas the problem of avoiding *uniformly unfair* systems (due to ill-founded explanations) is receiving less attention, in spite of being no less important.

These considerations call for the need of identifying some basic formal requirements that explanations should satisfy in order to lead to (deserving) trustworthy as well as (uniformly) fair AI systems. Indeed, providing a formal counterpart to these high-level principles appears to be crucial in order to carry out the following activities in a well-founded and non-ambiguous way: defining methods for quality verification and assurance from a human-centric perspective, comparing the adequacy of different systems on a uniform basis, providing guidelines for system development. Universal and absolute notions of trustworthiness and fairness being elusive, if not utopical, we share the suggestion that “the point is not complete fairness, but the need to establish metrics and thresholds for fairness that ensure trust in AI systems” (Dignum, 2021).

In turn, the investigation of formal requirements for explanations can benefit from a reference conceptual environment where their definition can be put in relation with some general foundational notions, whose suitability with respect to the human-centric perspective is well-established. Formal argumentation is an ideal candidate in this respect, for the reasons extensively illustrated in particular in Sections 3.1, 4.1 of Kakas et al. (2022) from which we limit ourselves to cite the emblematic statement that “Argumentation has a natural link to explanation.” Thus, it is not surprising that several works have focused on the use of argumentative techniques for a variety of explanation purposes (Cyras et al., 2021; Vassiliades et al., 2021). However, the study of argumentation-inspired formal properties related to human-centric issues like trustworthiness and fairness appears to have received lesser attention.

As a contribution to fill this gap, in this paper we use argumentation as a basis to formalize the property of *descriptive accuracy* (DA) described by Murdoch et al. (2019), for machine learning in general, as “the degree to which an interpretation method objectively captures the relationships learned by machine learning models.” DA appears to be a crucial requirement for any explanation: its absence would lead to the risk of misleading (if not deceptive) indications for the user (thus affecting trust and fairness). As such, one would expect that any explanation method is either able to enforce DA by construction or is equipped with a way to unearth possible violations of this fundamental property.

Specifically, we address the issue of defining argumentation-inspired formal counterparts (from simpler to more articulated) for the general notion of DA. In particular, our proposal leverages on two main sources from the argumentation literature: *Toulmin’s argument model* (Toulmin, 1958) and the formalism of *bipolar argumentation frameworks* (Cayrol and Lagasque-Schiex, 2005;

Amgoud et al., 2008; Cayrol and Lagasque-Schiex, 2013). In a nutshell, Toulmin’s argument model focuses on providing patterns for analyzing argument structure at a conceptual level. The most fundamental argument structure consists of three elements: claim, data and warrant. The *claim* of an argument is the conclusion it brings forward; the *data* provide evidence and facts which are the grounds in support of the claim; and the *warrant*, which could be implicit, links the data to the claim. Bipolar argumentation frameworks belong to the family of abstract argumentation formalisms pioneered by Dung (1995), where arguments are seen as abstract entities, and the main focus is on the relations among arguments, their meaning, and their role in the assessment of argument status. In particular, bipolar argumentation encompasses the basic relations of *attack* and *support* which provide a synthetic and powerful abstraction of the main kinds of dialectical interactions that may occur between two entities (see, for instance, Tversky and Kahneman (1992) and Dubois et al. (2008) for general analyses emphasizing the role of bipolarity in human decisions). A bipolar argumentation framework is hence a triple $(Args, Att, Supp)$ where $Att, Supp \subseteq Args \times Args$.

We will see that some of our abstractions for explanations can be put in correspondence with Toulmin’s model with an implicit warrant, whereas others can be seen as bipolar argumentation frameworks. Argumentation frameworks are typically equipped with “semantics” (e.g., notions of extensions) that may satisfy desirable properties: we define notions of DA drawing inspiration from some of these properties.

On these bases, focusing on the setting of *probabilistic classifiers*, we make the following contributions.

- We introduce three formal notions of DA (Section 5): *naive* DA, as a precursor to *dialectical* DA, both applicable to any probabilistic classifier, and *structural* DA, applicable to probabilistic classifiers that are equipped with a *structural description*, as is the case for *Bayesian network classifiers* (BCs) (see Bielza and Larrañaga, 2014 for an overview) and *Chain Classifiers* (CCs), resulting from chaining probabilistic classifiers (e.g., as is done for BCs by Read et al., 2009 and for other types of probabilistic classifiers by Blazek and Lin, 2021). These notions of DA are defined for generic abstractions of explanations, so that they can be applied widely to a variety of concrete notions instantiating the abstractions.
- We study whether concrete explanation methods (instantiating our abstract notions of explanation) satisfy our notions of DA (Section 6). We focus our analysis on (i) existing *feature attribution methods* from the literature, namely LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), as well as (ii) novel variants thereof and (iii) a novel method we define (which we refer to in short as DARX). We prove that: the methods (i) are not guaranteed to satisfy any of the formulations of DA we define; the variants (ii) are guaranteed to satisfy structural DA (by construction) but may still violate (naive and) dialectical DA; the DARX method is guaranteed to satisfy all the considered forms of DA (by construction), thus providing a proof of concept that our forms of DA are indeed satisfiable.

- We evaluate our forms of DA empirically on a variety of BCs and CCs (Section 7),¹ showing that they are often violated in practice by methods (i) and (ii).
- We describe a user study we conducted to gauge the importance of dialectical DA (our most “demanding” form of DA, applicable to any probabilistic classifier) to humans, when using explanations of probabilistic classifiers (Section 8), showing that this property predominantly aligns with human judgement.

3. Related work

A multitude of methods for providing explanations have been proposed (e.g., see the survey by Guidotti et al., 2019) and their desirable properties have been considered from a variety of perspectives (e.g., see the survey by Sokol and Flach, 2020). We draw inspiration from Murdoch et al. (2019) and focus, in particular, on their property of *descriptive accuracy* (DA) for (model-based or post-hoc) interpretable machine learning. As mentioned in the introduction, DA concerns the degree to which an interpretation (in our setting, explanation) method objectively captures the behavior of the machine-learned models. We will build on argumentative notions to provide three *formal* characterisations for DA, allowing evaluation of explanation methods for satisfaction of DA in precise terms.

DA is seen, in Murdoch et al. (2019), as a crucial property for achieving interpretable machine learning, alongside, in particular, *predictive accuracy*, wrt (test) data, of the predictions produced by the interpretations/explanations. Whereas DA is concerned with the inner workings of models, predictive accuracy is concerned with the input-output behavior thereof. Predictive accuracy is thus closely linked with properties of *fidelity* or *faithfulness* which have been considered by several works. For instance, in Guidotti et al. (2019) fidelity is defined as the capability of an explanation model to “accurately imitate a black-box predictor” and is measured in terms of accuracy score, F1-score, and so on, but wrt synthetic data capturing the behavior of the black-box. Analogously, in Lakkaraju et al. (2019), fidelity concerns the ability of an explanation to “faithfully mimic the behavior” of a model and is assessed in terms of the disagreement between the labels predicted by the explanation and the labels predicted by the model. In the case of explanations concerning a single instance, *local fidelity* has been defined as a measure of how well an explanation model approximates the original model in a neighborhood of the considered instance in need of explaining (Ribeiro et al., 2016; Alvarez-Melis and Jaakkola, 2017). In a similar vein, White and d’Avila Garcez (2020) define *counterfactual fidelity error* as the difference between the actual perturbation of a parameter needed to change the outcome in the original model and an estimate of that value, calculated using an approximate model.

¹ Note that some of these probabilistic classifiers are based on models which are, in principle, interpretable, like Bayesian networks. However we remark that interpretable models may still need explanations (Lipton, 2018; Ciatto et al., 2020; Du et al., 2020), e.g., because their size is beyond human interpretation capabilities or because lay users cannot understand probabilities.

Du et al. (2019) propose a *post-hoc* attribution method to explain the predictions of recurrent neural networks (RNNs) in text mining tasks with the goal of producing explanations, both at word-level and phrase-level which are faithful to the original RNN model. The method is specifically tailored to RNNs’ architecture as it resorts to computations on hidden state vectors. Faithfulness is evaluated empirically by computing a score based on the following idea: if one deletes the sentence with the highest attribution for a given prediction, one should then observe a significant drop in the probability of the predicted outcome, if the method is faithful. Thus, this work does not introduce a formal notion of faithfulness which is directly comparable to our characterization of descriptive accuracy and, in fact, the faithfulness score proposed is only indirectly related to the internal behavior of the RNN or of any other classifier.

The work by Adebayo et al. (2018) focuses on saliency methods used to highlight relevant features in images and shows that some of these methods are independent of both the data the model was trained on, and the model parameters, thus pointing out a lack of descriptive accuracy. Interestingly, but not completely surprisingly, it is shown that visual inspection of saliency maps may be misleading and some systematic tests (called sanity checks) are applied to verify whether the explanations depend on the data and the model parameters. The very interesting analysis carried out in this work provides striking evidence that the notion of descriptive accuracy requires more attention, while, differently from our present work, it does not include a proposal for an explicit formalization of this notion.

Yeh et al. (2019) address the problem of defining objective measures to assess explanations and propose, in particular, an infidelity measure, which can be roughly described as the difference between the effect of an input perturbation on the explanation and its effect on the output, and a sensitivity measure capturing the degree to which the explanation is affected by insignificant perturbations. Both measures use the classifier as a black box and hence there are no a priori guarantees about their ability to satisfy descriptive accuracy, as discussed in the present paper. Indeed, the authors apply a sanity check in the spirit of Adebayo et al. (2018) to verify whether the explanations generated to optimize the proposed measures are related to the model.

In the context of deep networks, Sundararajan et al. (2017) propose two axioms called Sensitivity and Implementation Invariance. The former consists of two requirements: (a) for every input and baseline that differ in one feature but have different predictions then the differing feature should be given a non-zero attribution; (b) if the function implemented by the deep network does not depend (mathematically) on some variable, then the attribution to that variable is always zero. The latter states that attributions should be identical for two functionally equivalent networks, where two networks are functionally equivalent if their outputs are equal for all inputs, despite having very different implementations. Sensitivity bears some similarity with the weakest notion we consider, namely naive descriptive accuracy, as they both refer to the role of individual variables and to ensure their relevance when present in explanations. However the perspective is slightly different as we essentially require that the presence of a feature in the explanation is somehow justified by the model, while Sundararajan et al. (2017) require that a feature is

present in the explanation under some specific conditions. Bridging these perspectives is an interesting issue for future work. The requirement of Implementation Invariance is motivated by the authors with the claim that attribution can be colloquially defined as assigning the blame (or credit) for the output to the input features. Such a definition does not refer to implementation details. While referring to the special (and rather unlikely in practice) situation where two internally different classifiers produce exactly the same output for the same input, we regard this requirement, which is somehow in contrast with descriptive accuracy, as partly questionable. Indeed, the fact that internal differences are reflected in the explanations may be, at least in principle, useful for some purposes like model debugging. If the differences concern the use of actually irrelevant features, we argue that this aspect should be captured by more general relevance-related criteria.

Chan et al. (2022) carry out a comparative study of faithfulness metrics for model interpretability methods in the context of natural language processing (NLP). Six faithfulness metrics are examined, all of which are based, with different nuances, on an evaluation of the role of the most important tokens in the classified sentences, in particular the common idea underlying these metrics is to compare the output of the classifiers for the same input with or without the most important tokens. These metrics use classifiers as black boxes and do not take into consideration their actual internal operation, so, while sharing the general goal of avoiding explanations that have loose correspondence with the explained model, their scope is somehow orthogonal to ours. Chan et al. (2022) observe that, though referring to the same basic principle, these metrics may provide contradictory outcomes, so that the most faithful method according to a metric is the worst with respect to another one. To address this problem, the authors propose a property of Diagnosticity, which refers to the capability of a metric to discriminate a more faithful interpretation from an unfaithful one (where, in practice, randomly generated interpretations are used as instances of unfaithful ones). Applying a possibly adapted notion of Diagnosticity in the context of our proposal appears an interesting direction of future work.

Mollas et al. (2022) propose *Altruist*, an approach for transforming the output of feature attribution methods into explanations using argumentation based on classical logic. In particular, *Altruist* is able to distinguish truthful vs. untruthful parts in a feature attribution and can work as a meta-explanation technique on top of a set of feature attribution methods. Similarly to our proposal, Mollas et al. (2022) assume that the importance weights produced by feature attribution methods are typically associated with a monotonic notion, and that end-users expect monotonic behavior when altering the value of some feature. On this basis, *Altruist* includes a module which assesses the truthfulness of an importance value by comparing the expected changes of the output, given some perturbations, with respect to the actual ones and then building an argumentation framework which is based on the predicates corresponding to the results of these comparisons and can be used to support a dialogue with the final user. The notion of truthfulness used in Mollas et al. (2022) refers to correspondence with users expectations rather than with internal model behavior and is thus complementary to our notion of descriptive accuracy. As both aspects are important

in practice, bridging them and investigating their relationships is a very interesting direction of future work. Also, the uses of argumentation in the two works are somehow complementary: while we resort to argumentation concepts as foundational notions, in Mollas et al. (2022) logic-based argumentation frameworks are used to support reasoning and dialogues about truthfulness evaluations.

Overall, whereas formal counterparts of predictive accuracy/faithfulness/fidelity have been extensively studied in the XAI literature, to the best of our knowledge, formal counterparts of DA appear to have received limited attention up to now. This gap is particularly significant for the classes of *post-hoc* explanations methods which, *per se*, have no relations with the underlying operation of the explained model and therefore cannot rely on any implicit assumption that DA is guaranteed, in a sense, by construction. This applies, in particular, to the family of *model-agnostic local explanation* methods, namely methods which are designed to be applicable to any model (and hence need to treat the model itself purely as a black-box) and whose explanations are restricted to illustrate individually a single outcome of the model without aiming to describe its behavior in more general terms. This family includes the well-known class of *additive feature attribution* methods, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), where the explanation for the outcome of a model basically consists in ascribing to each input feature a numerical weight. We will study our formalisations of DA in the context of both LIME and SHAP.

SHAP has been shown to be the only additive attribution method able to jointly satisfy three formal properties, called *local accuracy*, *missingness*, and *consistency* (see Lundberg and Lee, 2017 for details). These properties do not directly concern the internal working of the model and thus cannot be seen as forms of DA. Indeed, our analysis will show that SHAP, as well as LIME, are not guaranteed to satisfy our notions of DA— thus local accuracy, missingness and consistency do not suffice to enforce DA in our sense.

A variety of approaches devoted in particular to the explanation of Bayesian networks exist in the literature (Lacave and Díez, 2002; Mihaljevic et al., 2021). At a high level these approaches can be partitioned into three main families (Lacave and Díez, 2002): explanation of evidence (which concerns explaining observations by abducting the value of some unobserved variables), explanation of model (which aims at presenting the entire underlying model to the user), and explanation of reasoning. Explanation of reasoning is the one that best lends itself to fulfilling DA. According to Lacave and Díez (2002), it is in turn divided into: (i) explanation of the results obtained by the system and the reasoning process that produced them; (ii) explanation of the results not obtained by the system, despite the user's expectations; (iii) hypothetical reasoning, i.e., what results the system would have returned if one or more given variables had taken on different values from those observed. Our DARX approach is mainly related to point (i), even if it may support some form of hypothetical reasoning too. We remark that the spirit of DARX is not advancing the state of the art in explanations for Bayesian networks but rather providing a concrete example of a method satisfying the DA properties we introduce and showing that even with this baseline approach we can get

improvements with respect to popular model-agnostic methods, as concerns satisfaction of DA.

To introduce formal notions of DA we take inspiration from basic concepts in formal argumentation. As pointed out by [Cyras et al. \(2021\)](#), many popular methods for generating explanations in AI can be seen as implicitly argumentative, in addition to the vast literature on overtly argumentative approaches. These include the use of a variety of argumentation frameworks for explanation purposes, as surveyed in particular by [Cyras et al. \(2021\)](#), with a broad set of application domains, ranging from law and medical informatics to robotics and security, as discussed by [Vassiliades et al. \(2021\)](#). As pointed out by [Cyras et al. \(2021\)](#), however, in the literature the study of formal properties of argumentation frameworks have received by far more attention than the investigation of desirable properties of explanations, and our use of argumentation to drive the definition of explanation requirements, rather than of the explanation methods themselves, appears to be a novel investigation line to the best of our knowledge.

Last but not least, the human-centric approach requires that users' perspectives lie at the heart of the evaluation of AI explanation methods. Some works have identified properties naturally amenable to being assessed with humans, for example, [Murdoch et al. \(2019\)](#) propose *relevancy*, concerning the ability to provide insight for a particular audience into a chosen domain problem. It is widely acknowledged though that more user testing would be beneficial for evaluating XAI methods (e.g., see [Keane et al., 2021](#)). We contribute to this line of research by conducting a user study to assess whether our dialectical DA is in line with user expectations.

4. Preliminaries on probabilistic classifiers

As DA is inherently related to the internal operation of a model, rather than just to its input/output behavior, any formal notion of DA cannot be completely model-agnostic. It follows that an investigation of DA needs to find a balance between the obvious need of wide applicability and the potential advantages of model-tailored definitions. For this reason we will focus on the broad family of *probabilistic classifiers*.

We consider (discrete) probabilistic classifiers with *feature variables* $\mathbf{X} = \{X_1, \dots, X_m\}$ ($m > 1$) and *class variables* $\mathbf{C} = \{C_1, \dots, C_n\}$ ($n \geq 1$). Each (random) variable $V_i \in \mathbf{V} = \mathbf{X} \cup \mathbf{C}$ is equipped with a discrete set of possible *values*² Ω_{V_i} : we define the *feature space* as $\mathcal{X} = \Omega_{X_1} \times \dots \times \Omega_{X_m}$ and the *class space* as $\mathcal{C} = \Omega_{C_1} \times \dots \times \Omega_{C_n}$. From now on, we call any vector $\mathbf{x} \in \mathcal{X}$ an *input* and denote as $\mathbf{x}(X_i)$ the value of feature X_i in \mathbf{x} . Given input \mathbf{x} , a *probabilistic classifier* \mathcal{PC} computes, for each class variable C_i and value $\omega \in \Omega_{C_i}$, the probability $P(C_i = \omega | \mathbf{x})$ that C_i takes value ω , given \mathbf{x} .³ We then refer to the *resulting value* for a class variable $C_i \in$

\mathbf{C} given input \mathbf{x} as $\mathcal{PC}(C_i | \mathbf{x}) = \operatorname{argmax}_{\omega \in \Omega_{C_i}} P(C_i = \omega | \mathbf{x})$. [Table 1](#) gives a probabilistic classifier for a (toy) financial setting where the values of class variables *problematic external event* and *drop in consumer confidence* are determined based on the feature variables *company share price trend*, *devaluation of currency*, *healthy housing market* and *negative breaking news cycle*. Here, for any variable $V_i \in \mathbf{V}$, $\Omega_{V_i} = \{+, -\}$.

For $X_i \in \mathbf{X}$, we will abuse notation as follows, to simplify some of the formal definitions later in the paper: $\mathcal{PC}(X_i | \mathbf{x}) = \mathbf{x}(X_i)$ (basically, the “resulting value” for a feature variable, given an input, is the value assigned to that variable in the input) and $P(X_i = \mathbf{x}(X_i)) = 1$ (basically, the probability of a feature variable being assigned its value, in the given input, is 1). We will also use notation:

$$P(V=v|\mathbf{x}, \text{set}(V_i=v_i)) = \begin{cases} P(V=v|\mathbf{x}'), & \text{if } V_i \in \mathbf{X}, \\ P(V=v|\mathbf{x}, V_i=v_i), & \text{if } V_i \in \mathbf{C}, \end{cases}$$

where, in the first case, $\mathbf{x}'(V_i) = v_i$ and $\mathbf{x}'(V_j) = \mathbf{x}(V_j)$ for all $V_j \in \mathbf{X} \setminus \{V_i\}$. Basically, this notation allows to gauge the effects of changes in value for (input or class) variables on the probabilities computed by the classifiers (for assignments of values to any variables).

Various types of probabilistic classifiers exist. In [Section 7](#) we will experiment with (explanations for) a variety of (discrete) *Bayesian Classifiers* (BCs, see [Bielza and Larrañaga, 2014](#) for an overview), where the variables in \mathbf{V} constitute the nodes in a Bayesian network, i.e., a directed acyclic graph whose edges indicate probabilistic dependencies amongst the variables.⁴ We will also experiment with (explanations for) *chained probabilistic classifiers* (CCs, e.g., as defined by [Read et al. \(2009\)](#) for the case of BCs). These CCs result from the combination of simpler probabilistic classifiers (possibly, but not necessarily, BCs), using an ordering \succ_C over \mathbf{C} such that the value of any $C_i \in \mathbf{C}$ is treated as a feature value for determining the value of any $C_j \in \mathbf{C}$ with $C_j \succ_C C_i$, and thus a classifier computing values for C_i can be chained with one for computing values for C_j . For illustration, in [Table 2](#) we re-interpret the classifier from [Table 1](#) as a CC amounting to a chain of two classifiers, using $e \succ_C c$: the classifier (a) determines the value of c as an additional input for the classifier (b). Then, the overall classifier determines the value of c first based on the feature variables d , h and n , and then e based on s and c (treated as a feature variable in the chaining, thus implicitly taking into account d , h and n). Note that, in [Table 2](#) and throughout the paper, we abuse notation and use inputs for overall (chained) classifiers (\mathbf{x} in the caption of the table) as inputs of all simpler classifiers forming them (rather than the inputs' restriction to the specific input variables of the simpler classifiers).

For some families of probabilistic classifiers (e.g., for BCs) it is possible to provide a graphical representation which gives a synthetic view of the dependence and independence relations between the variables. In these cases, we will assume that the classifier is accompanied by a *structural description*, namely a set $\mathcal{SD} \subseteq \mathbf{V} \times \mathbf{V}$. The structural description identifies for each variable $V_j \in \mathbf{V}$ a (possibly empty) set of *parents* $\mathcal{PA}(V_j) = \{V_i \mid (V_i, V_j)\} \in \mathcal{SD}$ with the meaning that the evaluation of

² Thus, in this paper we discretise continuous variables, leaving their full treatment to future work.

³ Our focus is on explaining the outputs of classifiers, so we ignore how they are obtained, e.g., by hand or from data (subject to whichever bias), and how they perform computation.

⁴ BCs determine probabilities based on prior and conditional probabilities, e.g., using maximum a posteriori estimation. Given that our focus is on explaining, we ignore here how BCs are obtained.

TABLE 1 An example of probabilistic classifier with $X = \{s, d, h, n\}$ and $C = \{c, e\}$.

<i>s</i>	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−
<i>d</i>	+	+	+	+	−	−	−	−	+	+	+	+	−	−	−	−
<i>h</i>	+	+	−	−	+	+	−	−	+	+	−	−	+	+	−	−
<i>n</i>	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−
<i>c</i>	+	−	+	+	+	−	+	−	+	−	+	+	+	−	+	−
<i>P</i>	.60	.65	1	.60	.60	1	1	.65	.60	.65	1	.60	.60	1	1	.65
<i>e</i>	+	−	+	+	+	−	+	−	+	−	+	+	+	−	+	−
<i>P</i>	.60	1	.60	.60	.60	1	.60	1	1	.65	1	1	1	.65	1	.65

Here, e.g., for \mathbf{x} (highlighted in bold) such that $\mathbf{x}(s) = \mathbf{x}(d) = \mathbf{x}(h) = \mathbf{x}(n) = +$, $\mathcal{PC}(c|\mathbf{x}) = +$ (as $P(c = +|\mathbf{x}) = .60$), and $\mathcal{PC}(e|\mathbf{x}) = +$ (as $P(e = +|\mathbf{x}) = .60$).

TABLE 2 An example of chained probabilistic classifier (CC) with (a) the first probabilistic classifier \mathcal{PC}_1 with $X_1 = \{d, h, n\}$, $C_1 = \{c\}$, and (b) the second probabilistic classifier \mathcal{PC}_2 with $X_2 = \{s, c\}$, $C_2 = \{e\}$ (both inputs highlighted in bold).

(a)								
<i>d</i>	+	+	+	+	−	−	−	−
<i>h</i>	+	+	−	−	+	+	−	−
<i>n</i>	+	−	+	−	+	−	+	−
<i>c</i>	+	−	+	+	+	−	+	−
<i>P</i>	.60	.65	1	.60	.60	1	1	.65
(b)								
<i>s</i>	+	+	−	−	−	−	−	−
<i>c</i>	+	−	+	+	−	−	−	−
<i>e</i>	+	−	+	+	−	−	−	−
<i>P</i>	.60	1	1	1	.65	.65	.65	.65
(c)								

Here, e.g., for \mathbf{x} as in the caption of Table 1, $\mathcal{PC}(c|\mathbf{x}) = \mathcal{PC}_1(c|\mathbf{x}) = +$ and $\mathcal{PC}(e|\mathbf{x}) = \mathcal{PC}_2(e|\mathbf{x}, \text{set}(c = \mathcal{PC}_1(c|\mathbf{x}))) = +$. (c) A structural description for the CC in (a, b), shown as a graph.

V_j is completely determined by the evaluations of $\mathcal{PA}(V_j)$ in the classifier. In the case of BCs, the parents of each (class) variable correspond to the variables in its unique *Markov boundary* (Pearl, 1989; Neapolitan and Jiang, 2010) $\mathcal{M}: \mathbf{V} \rightarrow 2^{\mathbf{V}}$, where, for any $V_i \in \mathbf{V}$, $\mathcal{M}(V_i)$ is the \subseteq -minimal set of variables such that V_i is conditionally independent of all the other variables ($\mathbf{V} \setminus \mathcal{M}(V_i)$), given $\mathcal{M}(V_i)$. In the case of CCs, even when no information is available about the internal structure of the individual classifiers being chained, a structural description may be extracted to reflect the connections between features and classes. For illustration, for the CC in Tables 2a, b, the structural description is $\mathcal{SD} = \{(d, c), (h, c), (n, c), (s, e), (c, e)\}$, given in Table 2c as a graph.

We remark that notions similar to structural descriptions have been considered earlier in the literature. For instance, in Timmer et al. (2015) the argumentative notion of a support graph derived from a Bayesian network has been considered. This support graph

however is built with reference to a given variable of interest and is meant to facilitate the construction of arguments which provide a sort of representation of the reasoning inside the network. In our case we provide a structural description which does not refer to a single variable of interest and is not used for building explanations but rather to verify whether they satisfy structural DA, as will be described later.

In the remainder, unless specified otherwise, we assume as given a probabilistic classifier \mathcal{PC} with feature variables \mathbf{X} and class variables \mathbf{C} , without making any assumptions.

5. Formalizing descriptive accuracy

We aim to define DA, using argumentative notions as a basis, in a way which is independent of any specific explanation method (but with a focus on the broad class of local explanations, and specifically feature attribution methods to obtain them).

At a very abstract level, an explanation, whatever its structure is, can be regarded as including a set of *explanation elements* which are provided by the explainer to the explainee in order to justify some system *outcome*. Relationships between explanations under this abstract understanding and argumentative notions can be drawn at different levels. According to a first basic interpretation, the main components of an explanation can be put in correspondence with the essential parts of *Toulmin's argument model* (Toulmin, 1958): the system outcome can be regarded as an argument *claim*, while the explanation elements are the *data* supporting the claim; claim and data are connected (implicitly) by a *warrant*, namely the assumption on which the validity of the link from the data to the claim relies. In a more articulated interpretation, one can consider the existence of distinct argumentative relations underlying the explanation. Specifically, as mentioned in Section 2, we will focus on the fundamental relations of *attack* and *support* encompassed in *bipolar argumentation frameworks* (Amgoud et al., 2008).

According to both interpretations, the property of DA can be understood as the requirement that the argumentative structure underlying the explanation has a correspondence in the system being explained, and hence can be regarded as accurate. In particular, in the basic interpretation, we regard an explanation as satisfying DA if a suitable warrant, linking the explanation elements with the outcome, can be identified in the behavior of the system, while in the more articulated interpretation we require that the

relations of attack and support correspond to the existence of suitable bipolar influences within the system.

In order to convert these high-level considerations into formal definitions for both argumentative interpretations, we will consider different abstractions of the notion of (local) explanation, able to encompass a broad range of existing notions in the literature as instances. The abstractions we define are based on the combinations of alternative choices along two dimensions. On one hand, we consider two basic elements that an explanation may refer to: (1) *input features*; (2) pairs of variables representing *relations* between variables. When only input features are used then the resulting explanations are flat/shallow, only describing input/output behavior, whereas the inclusion of relations potentially allows for deeper explanation structures. On the other hand, we assume that the basic elements inside an explanation can be: (a) regarded as an undifferentiated set (we call these elements *unsigned*, in contrast with (b)); (b) partitioned into two sets according to their *positive* or *negative* role in the explanation. The combinations (1)-(a) and (2)-(a) will correspond respectively to the abstract notions of *unipolar* and *relational unipolar* explanations while the combinations (1)-(b) and (2)-(b) will correspond respectively to the notions of *bipolar* and *relational bipolar* explanations.⁵

Driven by argumentative interpretations for these forms of explanations, in terms of Toulmin's argument model and bipolar argumentation as highlighted above, we will introduce a notion of *naive* DA for all the kinds of abstract explanations we consider and a notion of *dialectical* DA tailored to the two cases of relational explanations. We see naive DA as a very weak pre-requisite for explanations, and prove that it is implied by dialectical DA for both bipolar and relational bipolar explanations (Propositions 1 and 2, resp.): thus, naive DA can be seen as a step toward defining dialectical DA. (Naive and) Dialectical DA are applicable to *any* probabilistic classifiers. In the specific setting of classifiers with underlying graph structures, such as BCs and CCs, we will also define a notion of *structural* DA for relational unipolar/bipolar explanations. Table 3 summarizes the definitions from this section, given below.

5.1. Unipolar explanations and naive DA

We begin with a very general notion of *unipolar explanation*: we only assume that, whatever the nature and structure of the explanation, it can be regarded at an abstract level as a *set of features*:

Definition 1. Given an input $\mathbf{x} \in \mathcal{X}$ and the resulting value $\omega = \mathcal{PC}(C|\mathbf{x})$ for class $C \in \mathcal{C}$ given \mathbf{x} , a *unipolar explanation* (for $C = \omega$, given \mathbf{x}) is a triple $\langle \mathbf{F}, C, \mathbf{x} \rangle$ where $\mathbf{F} \subseteq \mathbf{X}$.

⁵ We stress that what we call here explanations are in fact abstractions of what full-fledged explanations typically are (e.g., feature attribution methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) include, in addition to positive and negative features, a numerical value therefor). In this sense, we could more appropriately refer to our abstractions as *explanation skeletons*, but refrain to do so for simplicity of exposition.

It is easy to see that it is straightforward to derive unipolar explanations from the outcomes produced by existing explanation methods when they return features accompanied by additional information (e.g., feature importance as in the case of the attribution methods LIME and SHAP): basically, in these settings the unipolar explanations disregard the additional information, and amount to (a subset of) the set of features alone (e.g., the k most important features).

From an argumentative perspective, the features in a unipolar explanation can be regarded as the grounds (somewhat in Toulmin's sense) for justifying the resulting value assigned by the classifier to a class variable, for the input under consideration. Accordingly, we require that some form of warrant justifying the link of these grounds with the resulting value can be identified. This corresponds to the simplest form of DA, i.e., *naive DA*, whose intuition is that the features included in a unipolar explanation should be “relevant,” i.e., should play a role in the underlying model, as formally defined in the following.

Property 1. A unipolar explanation $\langle \mathbf{F}, C, \mathbf{x} \rangle$ satisfies *naive descriptive accuracy* iff for every $X_i \in \mathbf{F}$ there exists an input $\mathbf{x}' \in \mathcal{X}$ with $\mathbf{x}'(X_j) = \mathbf{x}(X_j)$ for every $X_j \neq X_i$ and with $\mathbf{x}'(X_i) \neq \mathbf{x}(X_i)$, such that, letting $\omega = \mathcal{PC}(C|\mathbf{x})$, it holds that $P(C = \omega|\mathbf{x}) \neq P(C = \omega|\mathbf{x}')$.

Naive DA holds when, for each individual feature, there is at least one case (i.e., an alternative input \mathbf{x}' to the input \mathbf{x} being explained) where a change in the value of the feature has an effect on the probability of the value of the class variable: thus, it is a rather weak requirement as it excludes individually “irrelevant” features from playing a role in the explanation. Note that this property can also be interpreted as a rudimentary form of *counterfactual reasoning* (of the form “what happens when the value of some variable changes?”). However, it is too weak to define counterfactual explanations (e.g., as first modeled in Tolomei et al., 2017; Wachter et al., 2017). Indeed, changes in probabilities, as in naive DA, may not lead to changes in classification, as required when defining counterfactual explanations. Furthermore, the notion of naive DA disregards considerations of “actionability” for counterfactual explanations, e.g., as addressed by Karimi et al. (2021). We leave formalization of DA for counterfactual explanations to future work.

For illustration, given the probabilistic classifier in Table 1 and \mathbf{x} as in the table's caption, the unipolar explanation $\langle \{s, d, h, n\}, c, \mathbf{x} \rangle$ does not satisfy naive DA, given that both s and d are “irrelevant” here: changing the value of either does not affect the probability of c . Instead, it is easy to see that $\langle \{h, n\}, c, \mathbf{x} \rangle$ satisfies naive DA.

5.2. Bipolar explanations and dialectical DA

Unipolar explanations consist of “minimal” information, i.e., just the features playing a role in explanations. At a finer level of granularity, corresponding to a greater degree of articulated argumentative interpretation, we consider *bipolar explanations*, where the features are partitioned into two sets: those having a positive, or *supporting*, effect on the resulting value and those having a negative, or *attacking*, effect. The notions of positive and negative effect may admit different specific interpretations in different contexts, the general underlying intuition being that

TABLE 3 Explanations with the characteristics they hold (as combinations of (1)-(2) and (a)-(b)) represented by ✓ and their DA properties (in italics) represented by ★.

	Unip. (Section 5.1)	Rel. Unip. (Section 5.2)	Bip. (Section 5.3)	Rel. Bip. (Section 5.4)
(1) input features	✓		✓	
(2) relations		✓		✓
(a) unsigned	✓	✓		
(b) positive or negative			✓	✓
<i>Basic DA</i>	★	★	★	★
<i>Dialectical DA</i>			★	★
<i>Structural DA</i>		★		★

the corresponding features provide, resp., reasons for and against the resulting value being explained. Whatever the interpretation, we assume that positive and negative features are disjoint, as a feature with a twofold role in an explanation could be confusing for the user.

Definition 2. Given an input $\mathbf{x} \in \mathcal{X}$ and the resulting value $\omega = \mathcal{PC}(C|\mathbf{x})$ for class $C \in \mathcal{C}$ given \mathbf{x} , a *bipolar explanation* (for $C = \omega$, given \mathbf{x}) is a quadruple $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$ where $\mathbf{F}_+ \subseteq \mathbf{X}$, $\mathbf{F}_- \subseteq \mathbf{X}$, and $\mathbf{F}_+ \cap \mathbf{F}_- = \emptyset$; we refer to features in \mathbf{F}_+ and \mathbf{F}_- resp. as *positive* and *negative reasons*.

It is easy to see that existing explanation methods can be regarded as producing bipolar explanations when those methods return features accompanied by additional positive or negative information (e.g., positive and negative feature importance as in the case of attribution methods such as LIME and SHAP): in these settings, as in the case of unipolar explanations, bipolar explanations disregard the additional information, and amount to (a subset of) the set of features with their polarity (e.g., the k features with the highest positive importance as positive features and the k features with the lowest negative importance as negative features).

Taking into account the distinction between positive and negative reasons, we introduce a property requiring that the dialectical role assigned to features is justified:

Property 2. A bipolar explanation $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$ satisfies *dialectical descriptive accuracy* iff for every $X_i \in \mathbf{F}_+ \cup \mathbf{F}_-$, for every $\mathbf{x}' \in \mathcal{X}$ with $\mathbf{x}'(X_j) = \mathbf{x}(X_j)$ for all $X_j \neq X_i$ and $\mathbf{x}'(X_i) \neq \mathbf{x}(X_i)$, letting $\omega = \mathcal{PC}(C|\mathbf{x})$, it holds that

- if $X_i \in \mathbf{F}_+$ then $P(C = \omega|\mathbf{x}) > P(C = \omega|\mathbf{x}')$;
- if $X_i \in \mathbf{F}_-$ then $P(C = \omega|\mathbf{x}) < P(C = \omega|\mathbf{x}')$.

In words, if a feature is identified as a positive (negative) reason for the resulting value for a class variable, given the input, the feature variable's value leads to increasing (decreasing, resp.) the posterior probability of the class variable's resulting value (with all other feature values unchanged). This has a direct correspondence with the properties of monotonicity considered in the literature for gradual argumentation semantics (Amgoud and Ben-Naim, 2018; Baroni et al., 2019) and we posit that this requirement ensures that each reason has a cognitively plausible dialectical meaning, faithful to human intuition, as we will examine in Section 8.

For illustration, in the running example with \mathcal{PC} in Table 1, the bipolar explanation $\langle \{d, n\}, \{h\}, c, \mathbf{x} \rangle$, given input \mathbf{x} as in the table's caption does not satisfy dialectical DA. Indeed, d is a positive reason in the explanation but, for \mathbf{x}' agreeing with \mathbf{x} on all features other than d (with $\mathbf{x}'(d) = -$), we obtain $P(c = +|\mathbf{x}) = .60 \not> P(c = +|\mathbf{x}') = .60$. Instead, it is easy to see that the bipolar explanation $\langle \{n\}, \{h\}, c, \mathbf{x} \rangle$, satisfies dialectical DA.

Note that the property of dialectical DA may not be satisfied by all re-interpretations of existing forms of explanations as bipolar explanations. As an example, consider *contrastive explanations* of the form proposed by Dhurandhar et al. (2018). Here, features are split into *pertinent positives and negatives*, which are those whose presence or absence, resp., is “relevant” to the resulting value being explained. If these pertinent positives and negatives are understood, resp., as positive and negative reasons in bipolar explanations, the latter do not satisfy dialectical DA, since both positive and negative pertinent features support the resulting value being explained. If, instead, pertinent positives and negatives are both understood as positive reasons, then the resulting bipolar explanations may satisfy dialectical DA: we leave the analysis of this aspect, and the definition of additional forms of DA e.g., able to distinguish between pertinent positives and negatives, for future work.

In general, unipolar explanations can be directly obtained from bipolar explanations by ignoring the distinction between positive and negative reasons, and the property of naive DA can be lifted:

Definition 3. A bipolar explanation $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$ satisfies *naive descriptive accuracy* iff the unipolar explanation $\langle \mathbf{F}_+ \cup \mathbf{F}_-, C, \mathbf{x} \rangle$ satisfies naive descriptive accuracy.

It is then easy to see that dialectical DA strengthens naive DA:⁶

Proposition 1. If a bipolar explanation $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$ satisfies dialectical DA then it satisfies naive DA.

5.3. Relational unipolar explanations and naive DA

Moving toward a notion of deeper explanations, we pursue the idea of providing a more detailed view of the relations between

⁶ All proofs not included in the paper can be found in Appendix 1.

variables of a probabilistic classifier, reflecting influences possibly occurring amongst them. To this purpose, we first introduce *relational unipolar explanations* as follows.

Definition 4. Given $\mathbf{x} \in \mathcal{X}$ and the resulting value $\omega = \mathcal{PC}(C|\mathbf{x})$ for $C \in \mathcal{C}$ given \mathbf{x} , a *relational unipolar explanation* (for $C = \omega$, given \mathbf{x}) is a triple $\langle \mathcal{R}, C, \mathbf{x} \rangle$ where $\mathcal{R} \subseteq \mathbf{V} \times \mathbf{V}$.

In words, a relational unipolar explanation includes a set \mathcal{R} of pairs of variables (i.e., a relation between variables) where $(V_i, V_j) \in \mathcal{R}$ indicates that the value of V_i has a role in determining the value of V_j , given the input.

For illustration, for \mathcal{PC} in Table 1, $\langle \{(s, e), (c, e)\}, e, \mathbf{x} \rangle$ may be a relational unipolar explanation for \mathbf{x} in the table's caption, indicating that s and c both influence (the value of) e . Note that relational unipolar explanations admit unipolar explanations as special instances: given a unipolar explanation $\langle \mathbf{F}, C, \mathbf{x} \rangle$, it is straightforward to see that $\langle \mathbf{F} \times \{C\}, C, \mathbf{x} \rangle$ is a relational unipolar explanation. However, as demonstrated in the illustration, relational unipolar explanations may include relations besides those between feature and class variables found in unipolar explanations. From an argumentative perspective, this corresponds to regarding the explanation as composed by a set of “finer grain” arguments, identifying not only the grounds for the explained outcome, but also for intermediate evaluations of the classifier, which in turn may provide grounds for the explained outcome and/or other intermediate evaluations.

The notion of naive DA can be naturally extended to relational unipolar explanations by requiring that a warrant based on relevance can be identified for each of the (implicit) finer arguments.

Property 3. A relational unipolar explanation $\langle \mathcal{R}, C, \mathbf{x} \rangle$ satisfies *naive descriptive accuracy* iff for every $(V_i, V_j) \in \mathcal{R}$, letting $v_i = \mathcal{PC}(V_i|\mathbf{x})$ and $v_j = \mathcal{PC}(V_j|\mathbf{x})$, there exists $v'_i \in \Omega_{V_i}$, $v'_i \neq v_i$, such that $P(V_j = v_j|\mathbf{x}) \neq P(V_j = v_j|\mathbf{x}, \text{set}(V_i = v'_i))$.

For illustration, for \mathcal{PC} in Table 1, $\langle \{(s, e), (n, e)\}, e, \mathbf{x} \rangle$ satisfies naive DA for \mathbf{x} in the table's caption, but $\langle \{(s, e), (d, e)\}, e, \mathbf{x} \rangle$ does not, as changing the value of d to $-$ (the only alternative value to $+$), the probability of $e = +$ remains unchanged.

It is easy to see that, for relational unipolar explanations $\langle \mathbf{F} \times \{C\}, C, \mathbf{x} \rangle$, corresponding to unipolar explanations $\langle \mathbf{F}, C, \mathbf{x} \rangle$, Property 1 is implied by Property 3.

5.4. Relational bipolar explanations and dialectical DA

Bipolarity can be directly enforced on relational explanations as follows.

Definition 5. Given an input $\mathbf{x} \in \mathcal{X}$ and the resulting value $\omega = \mathcal{PC}(C|\mathbf{x})$ for class $C \in \mathcal{C}$ given \mathbf{x} , a *relational bipolar explanation* (RX) is a quadruple $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$ where:

- $\mathcal{R}_+ \subseteq \mathbf{V} \times \mathbf{V}$, referred to as the set of *positive reasons*;
- $\mathcal{R}_- \subseteq \mathbf{V} \times \mathbf{V}$, referred to as the set of *negative reasons*;
- $\mathcal{R}_+ \cap \mathcal{R}_- = \emptyset$.

An RX can be seen as a graph of variables connected by edges identifying positive or negative reasons, i.e., as a bipolar argumentation framework (Cayrol and Lagasque-Schie, 2005). Here DA consists in requiring that the polarity of each edge is justified, which leads to the following definition, extending to relations the idea expressed in Property 2.

Property 4. An RX $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$ satisfies *dialectical descriptive accuracy* iff for every $(V_i, V_j) \in \mathcal{R}_+ \cup \mathcal{R}_-$, letting $v_i = \mathcal{PC}(V_i|\mathbf{x})$, $v_j = \mathcal{PC}(V_j|\mathbf{x})$, it holds that, for every $v'_i \in \Omega_{V_i} \setminus \{v_i\}$:

- if $(V_i, V_j) \in \mathcal{R}_+$ then $P(V_j = v_j|\mathbf{x}) > P(V_j = v_j|\mathbf{x}, \text{set}(V_i = v'_i))$;
- if $(V_i, V_j) \in \mathcal{R}_-$ then $P(V_j = v_j|\mathbf{x}) < P(V_j = v_j|\mathbf{x}, \text{set}(V_i = v'_i))$.

Similarly to dialectical descriptive accuracy for bipolar explanations, if, given the input, a variable V_i is categorized as a positive (negative) reason for the resulting value of another variable V_j , V_i 's value leads to increasing (decreasing, resp.) the posterior probability of V_j 's resulting value (with all values of the other variables playing a role in V_j 's value remaining unchanged).

Examples of RXs for the running example are shown as graphs in Figure 1 (where the nodes also indicate the values ascribed to the feature variables in the input \mathbf{x} and to the class variables by any of the toy classifiers in Tables 1, 2). Here, (iii) satisfies dialectical DA, since setting to $-$ the value of any variable with a positive (negative) reason to another variable will reduce (increase, resp.) the probability of the latter's value being $+$, whereas (ii) does not, since setting d to $-$ does not affect the probability of c 's value being $+$ and (i) does not since setting d to $-$ does not affect the probability of e 's value being $+$.

Similarly to the case of unipolar/bipolar explanations, relational unipolar explanations can be directly obtained from RXs by ignoring the distinction between positive and negative reasons, and the property of dialectical DA can be lifted:

Definition 6. An RX $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$ satisfies *naive descriptive accuracy* iff the relational unipolar explanation $\langle \mathcal{R}_+ \cup \mathcal{R}_-, C, \mathbf{x} \rangle$ satisfies naive descriptive accuracy.

It is then easy to see that dialectical DA strengthens naive DA:

Proposition 2. If an RX $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$ satisfies *dialectical DA* then it satisfies naive DA.

Note that bipolar explanations $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$ can be regarded as special cases of RXs, i.e., $\langle \{(X, C) | X \in \mathbf{F}_+\}, \{(X, C) | X \in \mathbf{F}_-\}, C, \mathbf{x} \rangle$ (indeed, the RX in Figure 1i is a bipolar explanation). Thus, from now on we will often refer to all forms of bipolar explanation as RXs.

5.5. Relational explanations and structural DA

When a classifier is equipped with a structural description, one can require that the relations used for explanation purposes in RXs are subsets of those specified by the structural description, so that the RXs correspond directly to (parts of) the inner working of the model. This leads to the following additional form of DA:

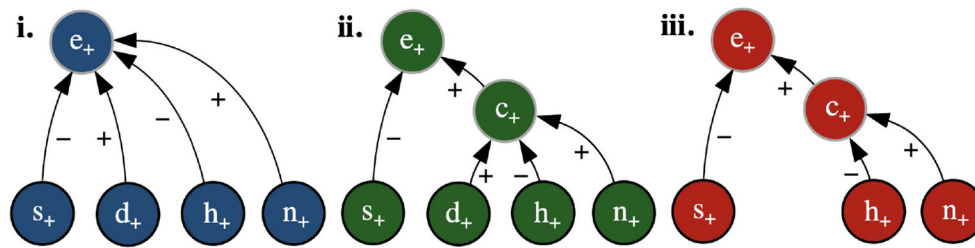


FIGURE 1

Example RXs (shown as graphs, with positive and negative reasons given by edges labeled + and -, resp.) with input \mathbf{x} such that $\mathbf{x}(s) = \mathbf{x}(d) = \mathbf{x}(h) = \mathbf{x}(n) = +$ (represented as s_+, d_+, h_+, n_+) and (resulting) class values $c = +$ (represented as c_+) and $e = +$ (represented as e_+).

Property 5. Given a probabilistic classifier \mathcal{PC} with structural description \mathcal{SD} :

- a relational unipolar explanation $\langle \mathcal{R}, C, \mathbf{x} \rangle$ satisfies *structural descriptive accuracy* iff $\mathcal{R} \subseteq \mathcal{SD}$; and
- an RX $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$ satisfies *structural descriptive accuracy* iff $\mathcal{R}_+ \cup \mathcal{R}_- \subseteq \mathcal{SD}$.

For instance, suppose that \mathcal{SD} is the structural description in Table 2c. Then, the RXs in Figures 1ii, iii satisfy structural DA, since all of the relations are contained within the structural description, while the RX in Figure 1i does not, since the relations from d, h and n to e are not present in the structural description.

6. Achieving descriptive accuracy in practice

In this section, we study the satisfaction of the proposed properties by explanation methods. We focus in particular on two existing methods in the literature, namely LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), and variants thereof that we design so that they satisfy structural DA. After showing that none of these methods satisfies all the properties introduced in Section 5, we introduce a novel form of explanation guaranteed to satisfy them, by definition. Thus, this novel form of explanation can be seen as a “champion” for our proposed forms of DA, showing that they can be satisfied in practice.

We start with LIME and SHAP. The explanations they produce (given an input \mathbf{x} and a classifier, computing $C = \omega$, given \mathbf{x}) basically consist in computing, for each feature $X_i \in \mathbf{X}$, a real number $w(\mathbf{x}, X_i, C)$ indicating the importance of X_i , which has assigned value $\mathbf{x}(X_i)$ in the given input \mathbf{x} , toward the probability of the class variable C being assigned value $\omega = \mathcal{PC}(C|\mathbf{x})$ by the classifier, in the context of \mathbf{x} .⁷ The absolute value of this number can be interpreted as a measure of the feature importance in the explanation, while its sign, in the context of explaining probabilistic

classifiers, indicates whether the feature has a positive or negative role wrt the classifier’s resulting value for the explained instance. Features which are assigned a value of zero can be regarded as irrelevant.⁸ Clearly, such explanations correspond to bipolar explanations $\langle \mathbf{F}_+, \mathbf{F}_-, C, \mathbf{x} \rangle$ as in Definition 2, with

- $\mathbf{F}_+ = \{X_i \in \mathbf{X} \mid w(\mathbf{x}, X_i, C) > 0\}$ and
- $\mathbf{F}_- = \{X_i \in \mathbf{X} \mid w(\mathbf{x}, X_i, C) < 0\}$.

In the remainder, with an abuse of terminology, we call these bipolar explanations *LIME/SHAP explanations*, depending on whether w is calculated using, resp., the method of LIME/SHAP. For illustration, consider the classifier in Table 1 and \mathbf{x} such that $\mathbf{x}(s) = \mathbf{x}(d) = \mathbf{x}(h) = \mathbf{x}(n) = +$, as in the caption of Figure 1, for which the classifier computes $e = +$. In this simple setting, SHAP computes $w(\mathbf{x}, s_+, e_+) = -0.20$, $w(\mathbf{x}, d_+, e_+) = 0.03$, $w(\mathbf{x}, h_+, e_+) = -0.05$, and $w(\mathbf{x}, n_+, e_+) = 0.25$ (adopting here the same conventions on variable assignments as in the caption of the Figure). This results in the SHAP explanation in Figure 1i. Thus features d and n (with their current values) are ascribed positive roles and s and h are ascribed negative roles in determining the outcome $\mathcal{PC}(e|\mathbf{x}) = +$. However, as stated earlier, for feature d this is in contrast with the property of naive DA. In fact, by inspection of Table 1, it can be noted that changing the value of this variable individually we would still have $P(e = +|\mathbf{x}) = 1$. To put it in intuitive terms, assigning a positive importance to this variable suggests to the user that its current value (namely $+$) has a role (though minor) in determining the outcome $e = +$, which is misleading. The following proposition generalizes these considerations:

⁸ While our property of naive DA requires that features included in an explanation are, in fact, relevant, a dual requirement would be that features not included in an explanation are, in fact, irrelevant. Addressing also this requirement corresponds to considering the implicit contents of an explanation too, i.e., all the features which are not presented to the user. However, the choice of features to be presented to the users may be determined also by their importance degree. In this sense, a feature which is not included does not necessarily need to be completely irrelevant, while a feature which is included definitely needs to be relevant. Based on this considerations, in this paper we focus on naive DA as a fundamental requirement, while we leave the investigation of more articulated versions of this property to future work.

⁷ We omit the formal definitions of how these well-known methods compute w , as the considerations in this paper on property satisfaction for LIME and SHAP are mostly based on empirical evaluation, supported by the standard implementations of LIME and SHAP, rather than their formal definition.

Proposition 3. In general, LIME and SHAP explanations are not guaranteed to satisfy naive nor dialectical DA.

The illustration above proves this result for SHAP explanations, by providing a counterexample to naive (and hence dialectical) DA in the context of the classifier in Table 1. The result for LIME explanations can be proved by introducing spurious features within trained probabilistic classifiers and showing that they play a role within LIME (see Appendix 1). As a side observation, in the appendix we also show empirically that approximate implementations of SHAP (the ones being used in practice, as an exact implementation of SHAP is practically unfeasible) also violate naive (and hence dialectical) DA.

Concerning structural DA, LIME and SHAP explanations may in general satisfy it only if $\mathbf{X} \times \mathbf{C} \subseteq \mathcal{SD}$, i.e., if the structural description includes all the possible relations from feature variables to class variables. This is, for instance, the case for naive BCs (Maron and Kuhns, 1960), but not for more general BCs or CCs. To overcome this limitation, generalizations of LIME and SHAP explanations can be defined so that they are guaranteed to satisfy structural DA by construction. This requires that the computation of (LIME/SHAP) w is applied not only to pairs with a feature and a class, but also, more generally, to any pairs of variables, following the underpinning structural description: in this way a bipolar argumentation framework satisfying structural DA is built.

Definition 7. Let \mathcal{PC} be a probabilistic classifier with structural description \mathcal{SD} . Given an input $\mathbf{x} \in \mathcal{X}$ and the resulting value $\omega = \mathcal{PC}(\mathbf{C}|\mathbf{x})$ for class $C \in \mathbf{C}$ given \mathbf{x} , a LIME/SHAP explanation satisfying structural DA (SDA-LIME/SDA-SHAP in short) is an RX $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$ such that $\mathcal{R}_+ \cup \mathcal{R}_- \subseteq \mathcal{SD}$ and

- $\mathcal{R}_+ = \{(V_i, V_j) \in \mathcal{SD} | w(\mathbf{x}, V_i, V_j) > 0\}$, and
- $\mathcal{R}_- = \{(V_i, V_j) \in \mathcal{SD} | w(\mathbf{x}, V_i, V_j) < 0\}$

where w is calculated, resp., using LIME/SHAP iteratively on the sub-classifiers induced by the structural description.

In practice, SDA-LIME and SDA-SHAP result from applying the attribution methods not on “black box” reasons (i.e., explaining class variables in terms of input features alone) but rather on reasons drawn from the structural description. In a nutshell, this amounts to applying LIME and SHAP by following the dependencies included in \mathcal{SD} , namely treating parents of class variables as features, in the context of sub-classifiers induced by \mathcal{SD} , step-wise. In the first iteration, for each class variable whose parents are all features (note that at least one such variable must exist), LIME and SHAP are applied to the sub-classifier consisting of the variable and its parents, and the weight computed for each parent is assigned to the link from the parent to the variable. Then, for the purposes of the subsequent iterations, each class variable to which this computation has been applied is marked as covered. As a consequence, new variables whose parents are all features or covered will be identified and LIME and SHAP will be applied to the relevant sub-classifiers as above. The process will terminate when reaching the coverage of all variables.

As a simple example, Figure 1ii gives an illustration of the application of SDA-SHAP for the structural description in Table 2c.

In the first iteration, SHAP is applied to the sub-classifier consisting of variable c (the only one whose parents are all features) and its parents, i.e., to the classifier in Table 2a, giving rise to $w(\mathbf{x}, d, c) = 0.04$, $w(\mathbf{x}, h, c) = -0.19$, $w(\mathbf{x}, n, c) = 0.18$. Then c is covered and SHAP can be applied to the classifier consisting of variable e and its parents (Table 2b), obtaining $w(\mathbf{x}, s, e) = -0.19$, $w(\mathbf{x}, c, e) = 0.31$ and completing the coverage of the variables.

Note that, like SDA-SHAP, Shapley Flow, recently proposed by Wang et al. (2021), generalizes SHAP so that reasons, rather than feature variables, are assigned a numerical weight. This is done using a causal model as the structural description for features and classes, in order to remove the risk that features not used by the model are assigned non-zero weights. Though featuring a similar high-level goal and sharing some basic idea, Shapley Flow significantly differs from SDA-SHAP. As a first remark, Shapley Flow is limited to single class variables, whereas SDA-SHAP can be used with probabilistic classifiers with any number of class variables. More importantly, in Shapley Flow the weights assigned to edges correspond to a notion of global flow rather than to a notion of importance of local influences, and hence have a different meaning wrt SDA-SHAP.

SDA-LIME and SDA-SHAP of course satisfy structural DA (by design) but fail to satisfy naive and dialectical DA.

Proposition 4. SDA-LIME & SDA-SHAP satisfy structural but are not guaranteed to satisfy naive nor dialectical DA.

The results above show that in order to guarantee the satisfaction of all the DA properties, an alternative approach to the construction of bipolar argumentation frameworks for explanation is needed. To this purpose, we introduce the novel *dialectically accurate relational explanations* (DARXs), whose definition is driven by the set of requirements we have identified.

Definition 8. Given a probabilistic classifier with structural description \mathcal{SD} , a *dialectically accurate relational explanation* (DARX) is a relational bipolar explanation $\langle \mathcal{R}_+, \mathcal{R}_-, C, \mathbf{x} \rangle$ where, letting $v_x = \mathcal{PC}(V_x|\mathbf{x})$ for any $V_x \in \mathbf{V}$:

- $\mathcal{R}_+ = \{(V_i, V_j) \in \mathcal{SD} | \forall v'_i \in \Omega_{V_i} \setminus \{v_i\} \text{ it holds that } P(V_j = v_j|\mathbf{x}) > P(V_j = v_j|\mathbf{x}, \text{set}(V_i = v'_i))\}$;
- $\mathcal{R}_- = \{(V_i, V_j) \in \mathcal{SD} | \forall v'_i \in \Omega_{V_i} \setminus \{v_i\} \text{ it holds that } P(V_j = v_j|\mathbf{x}) < P(V_j = v_j|\mathbf{x}, \text{set}(V_i = v'_i))\}$.

Proposition 5. DARXs are guaranteed to satisfy naive, structural and dialectical DA.

For illustration, suppose \mathcal{SD} corresponds exactly to the links in Figure 1iii. Then, this figure shows the DARX for e given the input in the figure's caption and the classifier in Table 1 (or Table 2). Here, the satisfaction of naive DA ensures that no spurious reasons, i.e., where the corresponding variables do not, in fact, influence one another, are included in the DARX. Note that, when explaining e with the same input, SHAP may draw a positive reason from d to e (as in Figure 1i) when, according to \mathcal{SD} , d does not directly affect e . Further, the satisfaction of dialectical DA means that each of the reasons in the DARX in Figure 1iii is guaranteed to have the desired dialectical effect (e.g., that the current value of n renders the (positive) prediction of c more likely, while the value of h has

the opposite effect). Meanwhile, the RXs (Figures 1i, ii) include the positive reasons from d , which have no bearing on either classification for this input.

Note that the bipolar argumentation frameworks representing DARXs are conceived as local explanations, i.e., they are meant to explain the behavior of the classifier given a specific input, not the behavior of the classifier in general. In other words, they assign a positive or negative role to variables with reference to the specific input considered and it may of course be the case that, given a different input, the same variable has a different role.

While DARX provides a notion of local explanation based on bipolar argumentation frameworks which is fully compliant with DA requirements, one may wonder whether its advantages are significant when applied to actual instances of probabilistic classifiers and whether it is viable in terms of performance. These questions are addressed by the empirical evaluation presented in next section.

7. Empirical evaluation

As mentioned in Section 4, we experiment with (chains of) BCs as well as chains (in the form of trees) of tree-based classifiers (referred to as *C-DTs* below). As far as BCs are concerned, we experiment with different types, corresponding to different restrictions on the structure of the underlying Bayesian network and conditional dependencies: naive BCs (*NBC*) (Maron and Kuhns, 1960); tree-augmented naive BCs (*TAN*) (Friedman et al., 1997); and *chains* of BCs (Zaragoza et al., 2011), specifically in the form of chains of the unrestricted BCs suggested in Provan and Singh (1995) (*CUBC*). We choose *C-DTs* and (chains of) BCs because they are naturally equipped with underlying structural descriptions, which allows us to evaluate structural DA, while they are popular methods with tabular data, e.g., in the case of BCs, for medical diagnosis (Lipovetsky, 2020; McLachlan et al., 2020; Stähli et al., 2021).⁹

Our experiments aim to evaluate the satisfaction/violation of structural and dialectical DA empirically for various concrete RXs (i.e., LIME, SHAP and their structural variants) when they are not guaranteed to satisfy the properties, as shown in Section 6.

The main questions we aim to address concern *actual DA* and *efficiency*, as follows. **Actual DA.** While some approaches may not be guaranteed to satisfy DA in general, they may for the most part in practice. *How much DA is achieved in the concrete settings of SHAP, LIME, SDA-SHAP and SDA-LIME explanations?* We checked the average percentages of reasons in LIME and SHAP explanations and in their structural counterparts which do not satisfy our notions of descriptive accuracy. The results are in Table 4. We

note that: (1) LIME often violates *naive descriptive accuracy*, e.g., in the *Child* and *Insurance* BCs, whereas SDA-LIME, SHAP and SDA-SHAP do not; (2) LIME and SHAP systematically violate *structural descriptive accuracy*; (3) LIME, SHAP and their structural counterparts often violate *dialectical descriptive accuracy*.

Efficiency. We have defined DARXs so that they are guaranteed to satisfy structural and dialectical DA. *Is the enforcement of these properties viable in practice, i.e., how expensive is it to compute DARXs?* Formally, the computational cost for DARXs can be obtained as follows. Let t_p be the time to compute a prediction and its associated posterior probabilities.¹⁰ The upper bound of the time complexity to compute a DARX is $T_{DARX}(\Omega) = O(t_p \cdot \sum_{V_i \in V} |\Omega_{V_i}|)$, which is *linear* with respect to the sum of the number of variables' values, making DARXs *efficient*.

8. Human experiment

Toward the goal of complying with human-centric requirements for explanations, we introduced *dialectical descriptive accuracy* as a cognitively plausible property supporting trust and fairness but lacking in some popular model-agnostic approaches. We hypothesize that dialectical DA aligns with human judgement. To assess our hypothesis, we conducted experiments on Amazon Mechanical Turk through a Qualtrics questionnaire with 72 participants. Of these, only 40 (56%) passed attention checks consisting of (1) basic questions for trivial information visualized on the screen and (2) timers checking whether the user was skipping very quickly through the questions. We used the Shuttle dataset to test our hypothesis. Indeed, this captures a setting with categorical (not only binary) observations, keeping participants' cognitive load low with an underlying classification problem easily understandable to lay users (see information about user expertise in Appendix 3).

We presented users with six questions, each accompanied by a DARX in the form exemplified in Figure 2 left, with six feature variables (i.e., *Wind Direction*, *Wind Strength*, *Positioning*, *Altimeter Error Sign*, *Altimeter Error Magnitude* and *Sky Condition*) assigned to various values and with corresponding predicted probability p for the (shown value of the) class variable *Recommended Control Mode*, as computed by our NBC for Shuttle (in Figure 2 left, $p = 0.979$ for the *Automatic* value of the class variable, given the values of the feature variables as shown). The graphical view demonstrated in the figure is a representation of a DARX as defined in Definition 8, with the green/red edges representing the positive/negative resp. reasons. We asked the users how they expected p to change when adding a positive (green arrow labeled with +) or negative (red arrow labeled with -) reason, e.g., Figure 2 shows how we asked the users what effect they thought that adding the positive reason *Altimeter Error Magnitude* would have (as in the DARX on the right). Specifically, we asked users to choose among options:

- (a) p increases;

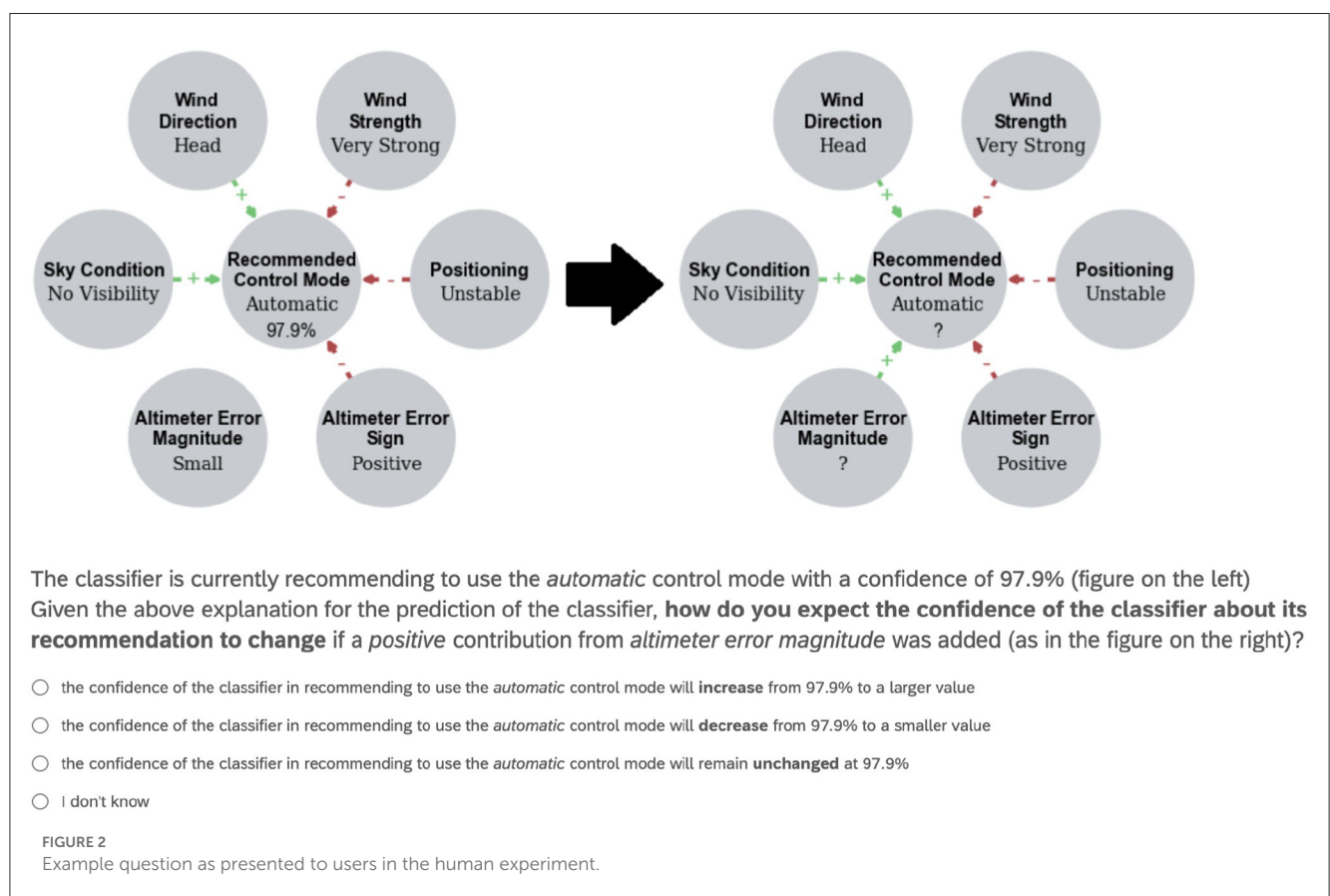
⁹ We use several datasets or (pre-computed) Bayesian networks, and deploy the best-performing type of the chosen type of classifier for each. Dataset sources were as follows: Shuttle (UCI, 2020), German (UCI, 2020), California (Kelley Pace and Barry, 1997), Child (BNlearn, 2020), Insurance (BNlearn, 2020), HELOC (Community, 2019) and LendingClub (LC) (Club, 2019). As structural descriptions, we use those described in Section 4. When training classifiers on datasets, we split them into train/test sets and optimize the hyper-parameters. See Appendix 2 for details on the datasets, training and performance, and for further details on the experiments.

¹⁰ In our experiments, using a machine with Intel i9-9900X at 3.5Ghz and 32GB of RAM with no GPU acceleration, t_p ranges from 3μs for the simplest NBC to 436ms for the most complex chain classifier.

TABLE 4 Average percentages of reasons (over 100 samples) violating DA (i.e., $\{ (V_i, V_j) \in \mathcal{R}_- \cup \mathcal{R}_+ \text{ such that } (V_i, V_j) \text{ violates DA} \} / |\mathcal{R}_- \cup \mathcal{R}_+|$) for several instantiated RXs.

Dataset	Classifier*	SHAP			LIME			SDA-SHAP		SDA-LIME	
		Naive	Structural	Dialectical	Naive	Structural	Dialectical	Naive	Dialectical	Naive	Dialectical
Shuttle	NBC	0%	0% [†]	16.43%	0%	0% [†]	17.14%	‡	‡	‡	‡
German	NBC	0%	0% [†]	54.56%	0%	0% [†]	49.55%	‡	‡	‡	‡
California	TAN	0%	0% [†]	16.75%	0%	0% [†]	16.75%	‡	‡	‡	‡
Insurance	CUBC	0%	67.07%	78.77%	59.56%	89.26%	93.07%	0%	41.77%	0%	42.56%
Child	CUBC	0%	70.97%	75.35%	63.74%	89.59%	91.16%	0%	21.18%	0%	21.18%
HELOC	C-DTs	51.77%	100%	94.42%	62.21%	100%	97.87%	25.60%	77.88%	31.21%	82.09%
LC	C-DTs	16.19%	100%	94.47%	72.95%	100%	98.57%	0%	52.26%	0%	57.63%

(*) NBC (Naive BC), TAN (Tree-Augmented NBC), CUBC (Chain of Unrestricted BCs), C-DTs (Chain of Decision Trees); (†) results must be 0.0% due to the BC type; (‡) SDA-LIME and SDA-SHAP explanations are equal to LIME and SHAP, resp., due to the BC type.



- (b) p decreases;
- (c) p remains unchanged; and
- (d) I don't know,

as indicated in Figure 2. For our hypothesis to hold we expected users to select answer (a) when adding positive reasons (as in Figure 3) and to select answer (b) when adding negative reasons. We also assessed how consistent users were, dividing the results based on the number of questions (out of 6) users answered following the same pattern, e.g., consistency of 6 means either

all answers aligned with our hypothesis or all answers did not, while consistency of 3 means half of the answers aligned and half did not.

The results are shown in Figure 3: here, when computing the p-values against the null hypothesis of random answers (50%-50%) we used the multinomial statistical test. We note that: (1) in all cases users' answers were predominantly in line with our expectations; and (2) participants that were consistent in answering more questions were more likely to agree with our hypothesis.

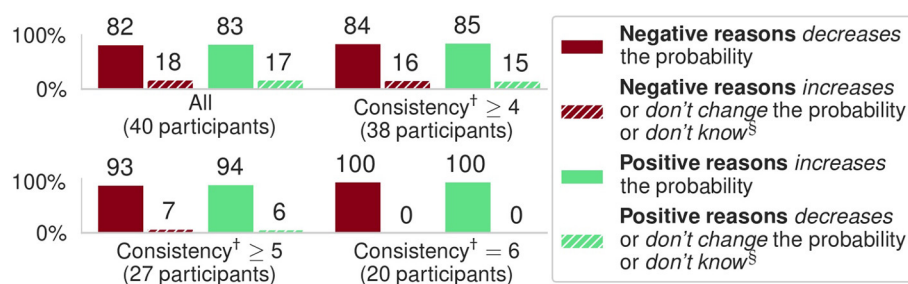


FIGURE 3

Results of the experiments with 40 participants; all results are significant ($p < 0.001$) against the null hypothesis of random answers. Here, "Negative/Positive reasons" refers to adding negative/positive contributions, resp., from features, as illustrated in Figure 2. (†) Consistency represents the number of questions (out of 6) users answered following the same pattern (also unexpected ones, e.g., that negative reasons increase probability). (§) We aggregated all results for unexpected answers in these bar plots.

9. Conclusions

In this paper we have studied how to define and enforce properties of explanations for the outputs of AI models (focusing on probabilistic classifiers), so that they can be deemed trustworthy and fair, in the sense that they do not mislead their users. Specifically, we have introduced a three-fold notion of DA for explanations of probabilistic classifiers, which, despite its intuitiveness, is often not satisfied by prominent explanation methods, and shown that it can be satisfied, by design, by the novel explanation concept of DARXs. We have performed a wide-ranging evaluation with theoretical results and experiments in a variety of data-centric settings and with humans wrt explanation baselines, highlighting the importance of our most demanding notion of DA (dialectical DA), from a human perspective. This demonstrates how DA, which has thus far been overlooked in the explainable AI literature, could be a critical component in achieving trustworthy and fair systems, in line with the principles of human-centric AI. We have built our definitions of DA and DARX around notions inspired by formal notions of argumentation, thus providing some instantiated evidence about the foundational role of argumentation for human-centric AI, on which the present special issue is focused.

Our work opens several avenues for future work. It would be interesting to experiment with other forms of probabilistic classifiers, including (chained) neural networks, possibly in combination with methods for extracting *causal models* from these classifiers (e.g., as in Kyono et al., 2020) to provide structural descriptions for satisfying structural DA. It would also be interesting to study the satisfaction of (suitable variants of) DA, e.g., those incorporating zero-valued variables as mentioned previously, by other forms of explanations, including *minimum cardinality* explanations (Shih et al., 2018) and set-based explanations (Dhurandhar et al., 2018; Ignatiev et al., 2019). We also note that our proposed methodology in this paper can support human users' full understandings of model intricacies toward leading to their outputs. However, as with other explanation models, there is a fine line between explainability and manipulability. Thus, it would be interesting to explore potential risks in revealing the inner workings of probabilistic

classifiers to end users, as this may empower users to manipulate them. We would also like to extend the human experiment described in Section 8 to present a more rigorous assessment of our notions of DA, e.g., assessing structural DA, which would require users who are able to appreciate the model's underpinning structure. Last but not least, while the human experiment provided encouraging indications about the cognitive plausibility of the proposed approach, the present research needs to be complemented by an investigation focused on the Human-Computer Interaction (HCI) aspects involved in properly conveying explanations to users. The fact that HCI principles and methodologies are of paramount importance in human-centric AI has been pointed out by several works in the literature (see e.g., Xu, 2019; Shneiderman, 2020), which also stress the need to properly take into account human and ethical factors. In particular, interactivity is a key factor to address the basic tension between interpretability and accuracy, especially when dealing with complex models (Weld and Bansal, 2019). This is demonstrated, for instance, in case studies where suitable interaction mechanisms are used to allow users to combine global and local explanation paradigms (Hohman et al., 2019) or to enable heuristic cooperation between users and machine in a challenging context like the analysis of complex data in the criminal justice domain (Lettieri et al., 2022).

Author's note

This paper extends (Albini et al., 2022) in various ways; in particular we introduce novel variants of LIME and SHAP, which satisfy structural DA by design, and we undertake a human experiment examining our approach along the metric of consistency.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

EA had the main responsibility for the implementation and experiments. All authors contributed equally to the conceptual analysis, formal development, and writing of the paper. All authors contributed to the article and approved the submitted version.

Funding

PB had been partially supported by the GNCS-INdAM project CUP_E55F22000270001. FT and AR were partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934) and by J. P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme.

References

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B. (2018). "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, eds S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett (Montreal, QC), 9525–9536.
- Albini, E., Rago, A., Baroni, P., and Toni, F. (2022). "Descriptive accuracy in explanations: The case of probabilistic classifiers," in *Scalable Uncertainty Management - 15th International Conference, SUM 2022 (Paris)*, 279–294.
- Alvarez-Melis, D. and Jaakkola, T. S. (2017). "A causal framework for explaining the predictions of black-box sequence-to-sequence models," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017 (Copenhagen)*, 412–421.
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., et al. (2019). "Guidelines for human-ai interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19 (New York, NY: Association for Computing Machinery)*, 1–13.
- Amgoud, L. and Ben-Naim, J. (2018). Evaluation of arguments in weighted bipolar graphs. *Int. J. Approx. Reason.* 99, 39–55. doi: 10.1016/j.ijar.2018.05.004
- Amgoud, L., Cayrol, C., Lagasque-Schiek, M., and Livet, P. (2008). On bipolarity in argumentation frameworks. *Int. J. Intell. Syst.* 23, 1062–1093. doi: 10.1002/int.20307
- Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., et al. (2017). Towards artificial argumentation. *AI Mag.* 38, 25–36. doi: 10.1609/aimag.v38i3.2704
- Baroni, P., Gabbay, D., Giacomin, M., and van der Torre, L. (Eds.). (2018). *Handbook of Formal Argumentation*. College Publications.
- Baroni, P., Rago, A., and Toni, F. (2019). From fine-grained properties to broad principles for gradual argumentation: a principled spectrum. *Int. J. Approx. Reason.* 105, 252–286. doi: 10.1016/j.ijar.2018.11.019
- Bielza, C. and Larra naga, P. (2014). Discrete bayesian network classifiers: a survey. *ACM Comput. Surv.* 47, 1–5. doi: 10.1145/2576868
- Binns, R. (2020). "On the apparent conflict between individual and group fairness," in *FAT* '20: Conference on Fairness, Accountability, and Transparency (Barcelona)*, 514–524.
- Blazek, P. J., and Lin, M. M. (2021). Explainable neural networks that simulate reasoning. *Nat. Comput. Sci.* 1, 607–618. doi: 10.1038/s43588-021-00132-w
- BNlearn (2020). *Bayesian Network Repository - an R Package for Bayesian Network Learning and Inference*.
- Cayrol, C., and Lagasque-Schiek, M. (2005). "On the acceptability of arguments in bipolar argumentation frameworks," in *Proceedings of the 8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (Barcelona)*, 378–389.
- Cayrol, C., and Lagasque-Schiek, M. (2013). Bipolarity in argumentation graphs: Towards a better understanding. *Int. J. Approx. Reason.* 54, 876–899. doi: 10.1016/j.ijar.2013.03.001
- Chan, C. S., Kong, H., and Liang, G. (2022). "A comparative study of faithfulness metrics for model interpretability methods," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, eds S. Muresan, P. Nakov, and A. Villavicencio (Dublin: Association for Computational Linguistics), 5029–5038.
- Ciatto, G., Calvaresi, D., Schumacher, M. I., and Omicini, A. (2020). "An abstract framework for agent-based explanations in AI," in *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20 (Auckland)*, 1816–1818.
- Club, L. (2019). *Lending Club Loans*.
- Community, F. (2019). *Explainable Machine Learning Challenge*.
- Cyras, K., Rago, A., Albini, E., Baroni, P., and Toni, F. (2021). "Argumentative XAI: a survey," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021 (Virtual Event/Montreal, QC)*, 4392–4399.
- Dhurandhar, A., Chen, P., Luss, R., Tu, C., Ting, P., Shanmugam, K., et al. (2018). "Explanations based on the missing: towards contrastive explanations with pertinent negatives," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018 (Montreal, QC)*, 590–601.
- Dignum, V. (2021). "The myth of complete AI-fairness," in *Artificial Intelligence in Medicine*, eds A. Tucker, P. Henriques Abreu, J. Cardoso, P. Pereira Rodrigues, and D. Ria no (Cham: Springer International Publishing), 3–8.
- Du, M., Liu, N., and Hu, X. (2020). Techniques for interpretable machine learning. *Commun. ACM*, 63, 68–77. doi: 10.1145/3359786
- Du, M., Liu, N., Yang, F., Ji, S., and Hu, X. (2019). "On attribution of recurrent neural network predictions via additive decomposition," in *The World Wide Web*

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

Any views or opinions expressed herein are solely those of the authors.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2023.1099407/full#supplementary-material>

- Conference, WWW 2019, eds L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia (San Francisco, CA: ACM), 383–393.
- Dubois, D., Fargier, H., and Bonnefon, J. (2008). On the qualitative comparison of decisions having positive and negative features. *J. Artif. Intell. Res.* 32, 385–417. doi: 10.1613/jair.2520
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77, 321–358. doi: 10.1016/0004-3702(94)00041-X
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012). “Fairness through awareness,” in *Innovations in Theoretical Computer Science* (Cambridge, MA), 214–226.
- Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Mach. Learn.* 29, 131–163. doi: 10.1023/A:1007465528199
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 1–93. doi: 10.1145/3236009
- Heidari, H., Loi, M., Gummad, K. P., and Krause, A. (2019). “A moral framework for understanding fair ML through economic models of equality of opportunity,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019* (Atlanta, GA), 181–190.
- Hohman, F., Head, A., Caruana, R., DeLine, R., and Drucker, S. M. (2019). “Gamut: a design probe to understand how data scientists understand machine learning models,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19* (New York, NY: Association for Computing Machinery), 1–13.
- Hutchinson, B., and Mitchell, M. (2019). “50 years of test (un)fairness: lessons for machine learning,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019* (Atlanta, GA), 49–58.
- Ignatiev, A., Narodytska, N., and Marques-Silva, J. (2019). “Abduction-based explanations for machine learning models,” in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* (Honolulu, HI), 1511–1519.
- Independent high-level expert group on Artificial Intelligence set up by the European Commission (2019). *Ethics guidelines for trustworthy AI*. Technical report.
- Jacovi, A., Marasovic, A., Miller, T., and Goldberg, Y. (2021). “Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI,” in *FACCT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event/Toronto, ON), 624–635.
- Kakas, A., Dietz, E., and Michael, L. (2022). Argumentation: A calculus for human-centric AI. *Front. Artif. Intell.* 5, 955579. doi: 10.3389/frai.2022.955579
- Karimi, A., Schölkopf, B., and Valera, I. (2021). “Algorithmic recourse: from counterfactual explanations to interventions,” in *FACCT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency* (Toronto, ON: ACM), 353–362.
- Keane, M. T., Kenny, E. M., Delaney, E., and Smyth, B. (2021). “If only we had better counterfactual explanations: five key deficits to rectify in the evaluation of counterfactual XAI techniques,” in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021*, (Montreal, QC; Virtual Event), 4466–4474.
- Kelley Pace, R. and Barry, R. (1997). Sparse spatial autoregressions. *Stat. Probab. Lett.* 33, 291–297. doi: 10.1016/S0167-7152(96)00140-X
- Kyono, T., Zhang, Y., and van der Schaar, M. (2020). “CASTLE: regularization via auxiliary causal graph discovery,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.
- Lacave, C., and Diez, F. J. (2002). A review of explanation methods for bayesian networks. *Knowl. Eng. Rev.* 17, 107–127. doi: 10.1017/S026988890200019X
- Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2019). “Faithful and customizable explanations of black box models,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019* (Honolulu, HI), 131–138.
- Lettieri, N., Guarino, A., Malandrino, D., and Zaccagnino, R. (2022). Knowledge mining and social dangerousness assessment in criminal justice: metaheuristic integration of machine learning and graph-based inference. *Artif. Intell. Law.* doi: 10.1007/s10506-022-09334-7
- Lipovetsky, S. (2020). Let the evidence speak - using Bayesian thinking in law, medicine, ecology and other areas. *Technometrics* 62, 137–138. doi: 10.1080/00401706.2019.1708677
- Lipton, Z. C. (2018). The myths of model interpretability. *Commun. ACM* 61, 36–43. doi: 10.1145/3233231
- Lundberg, S. M. and Lee, S. (2017). “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017* (Long Beach, CA), 4765–4774.
- Maron, M. E. and Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *J. ACM* 7, 216–244. doi: 10.1145/321033.321035
- McLachlan, S., Dube, K., Hitman, G. A., Fenton, N. E., and Kyrimi, E. (2020). Bayesian networks in healthcare: distribution by medical condition. *Artif. Intell. Med.* 107, 101912. doi: 10.1016/j.artmed.2020.101912
- Mihaljevic, B., Bielza, C., and Larra naga, P. (2021). Bayesian networks for interpretable machine learning and optimization. *Neurocomputing* 456, 648–665. doi: 10.1016/j.neucom.2021.01.138
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007
- Mollas, I., Bassiliades, N., and Tsoumakas, G. (2022). “Altruist: argumentative explanations through local interpretations of predictive models,” in *SETN 2022: 12th Hellenic Conference on Artificial Intelligence, Corfu Greece* (Greece: ACM), 1–21.
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. U.S.A.* 116, 22071–22080. doi: 10.1073/pnas.1900654116
- Neapolitan, R. E. and Jiang, X. (2010). *Probabilistic Methods for Financial and Marketing Informatics*. Elsevier.
- Pearl, J. (1989). *Probabilistic Reasoning in Intelligent Systems-Networks of Plausible Inference*. Morgan Kaufmann.
- Provan, G. M., and Singh, M. (1995). “Learning Bayesian networks using feature selection,” in *Learning from Data - Fifth International Workshop on Artificial Intelligence and Statistics, AISTATS 1995* (Key West, FL), 291–300.
- Rago, A., Cocarascu, O., Bechlivandis, C., Lagnado, D. A., and Toni, F. (2021). Argumentative explanations for interactive recommendations. *Artif. Intell.* 296, 103506. doi: 10.1016/j.artint.2021.103506
- Räz, T. (2021). “Group fairness: independence revisited,” in *FACCT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event/Toronto, ON), 129–137.
- Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2009). “Classifier chains for multi-label classification,” in *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled* (Slovenia), 254–269.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “Why should I trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA), 1135–1144.
- Shih, A., Choi, A., and Darwiche, A. (2018). “A symbolic approach to explaining Bayesian network classifiers,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018* (Stockholm), 5103–5111.
- Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Trans. Interact. Intell. Syst.* 10, 764. doi: 10.1145/3419764
- Sokol, K. and Flach, P. A. (2020). “Explainability fact sheets: a framework for systematic assessment of explainable approaches,” in *FAT* '20: Conference on Fairness, Accountability, and Transparency* (Barcelona), 56–67.
- Stähli, P., Frenz, M., and Jaeger, M. (2021). Bayesian approach for a robust speed-of-sound reconstruction using pulse-echo ultrasound. *IEEE Trans. Med. Imaging* 40, 457–467. doi: 10.1109/TMI.2020.3029286
- Sundararajan, M., Taly, A., and Yan, Q. (2017). “Axiomatic attribution for deep networks,” in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, eds D. Precup and Y. W. Teh (Sydney, NSW: PMLR), 3319–3328.
- Timmer, S., Meyer, J. C., Prakken, H., Renooij, S., and Verheij, B. (2015). “Explaining Bayesian networks using argumentation,” in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 13th European Conference ECSQARU 2015* (Compiegne), 83–92.
- Tolomei, G., Silvestri, F., Haines, A., and Lalmas, M. (2017). “Interpretable predictions of tree-based ensembles via actionable feature tweaking,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS), 465–474.
- Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.
- Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertain.* 5, 297–323. doi: 10.1007/BF00122574
- UCI, C. (2020). *Machine Learning Repository*.
- Vassiliades, A., Bassiliades, N., and Patkos, T. (2021). Argumentation and explainable artificial intelligence: a survey. *Knowl. Eng. Rev.* 36, e5. doi: 10.1017/S0269888921000011
- Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv. JL Tech.* 31, 841. doi: 10.2139/ssrn.3063289
- Wang, J., Wiens, J., and Lundberg, S. M. (2021). “Shapley flow: a graph-based approach to interpreting model predictions,” in *The 24th International Conference on*

Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021 (Virtual Event), 721–729.

Weld, D. S. and Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 70–79. doi: 10.1145/3282486

White, A. and d'Avila Garcez, A. S. (2020). “Measurable counterfactual local explanations for any classifier,” in *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)* (Santiago de Compostela), 2529–2535.

Xu, W. (2019). Toward human-centered ai: a perspective from human-computer interaction. *Interactions* 26, 42–46. doi: 10.1145/3328485

Yeh, C., Hsieh, C., Suggala, A. S., Inouye, D. I., and Ravikumar, P. (2019). “On the (in)fidelity and sensitivity of explanations,” in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019*, eds H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett (Vancouver, BC), 10965–10976.

Zaragoza, J. H., Sucar, L. E., Morales, E. F., Bielza, C., and Larra naga, P. (2011). “Bayesian chain classifiers for multidimensional classification,” in *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence* (Barcelona), 2192–2197.

Zerilli, J., Bhatt, U., and Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns* 3, 100455. doi: 10.1016/j.patter.2022.100455



OPEN ACCESS

EDITED BY
Emmanuelle Dietz,
Airbus, Germany

REVIEWED BY
Marcos Cramer,
Technical University Dresden, Germany
Bettina Fazzinga,
University of Calabria, Italy

*CORRESPONDENCE
Christian Straßer
✉ christian.strasser@rub.de

SPECIALTY SECTION
This article was submitted to
Machine Learning and Artificial Intelligence,
a section of the journal
Frontiers in Artificial Intelligence

RECEIVED 29 December 2022

ACCEPTED 20 March 2023

PUBLISHED 19 May 2023

CITATION
Straßer C and Michajlova L (2023) Evaluating
and selecting arguments in the context of
higher order uncertainty.
Front. Artif. Intell. 6:1133998.
doi: 10.3389/frai.2023.1133998

COPYRIGHT
© 2023 Straßer and Michajlova. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Evaluating and selecting arguments in the context of higher order uncertainty

Christian Straßer* and Lisa Michajlova

Institute for Philosophy II, Ruhr University Bochum, Bochum, Germany

Human and artificial reasoning has to deal with uncertain environments. Ideally, probabilistic information is available. However, sometimes probabilistic information may not be precise or it is missing entirely. In such cases we reason with higher-order uncertainty. Formal argumentation is one of the leading formal methods to model defeasible reasoning in artificial intelligence, in particular in the tradition of Dung's abstract argumentation. Also from the perspective of cognition, reasoning has been considered as argumentative and social in nature, for instance by Mercier and Sperber. In this paper we use formal argumentation to provide a framework for reasoning with higher-order uncertainty. Our approach builds strongly on Haenni's system of probabilistic argumentation, but enhances it in several ways. First, we integrate it with deductive argumentation, both in terms of the representation of arguments and attacks, and in terms of utilizing abstract argumentation semantics for selecting some out of a set of possibly conflicting arguments. We show how our system can be adjusted to perform well under the so-called rationality postulates of formal argumentation. Second, we provide several notions of argument strength which are studied both meta-theoretically and empirically. In this way the paper contributes a formal model of reasoning with higher-order uncertainty with possible applications in artificial intelligence and human cognition.

KEYWORDS

abstract argumentation, probabilistic argumentation, argument strength, higher-order uncertainty, reasoning with uncertainty, non-monotonic logic

1. Introduction

1.1. Reasoning with uncertainties

Many sources of information provide uncertain information. Such information may come with probabilistic estimations of how likely specific events are (think of a weather report), in which case we deal with (precise or first order) probabilistic uncertainty. However, often probabilistic information is missing, or the probabilities are not known precisely, in which case we deal with higher-order uncertainty (in short, HOU). HOU occurs when the underlying probability distribution is not or only partially known.¹ We illustrate the role of HOU with two examples.

¹ We note that following Keynes and Knight "uncertainty" is often used for non-probabilistic uncertainties in contradistinction to "risk" (which in this paper is first order uncertainty). For a discussion on the subtle differences in Knight and Keynes and for further discussion on the pairs of distinctions risk-vs-uncertainty and the related probabilistic-vs-non-probabilistic uncertainty (see O'Donnell, 2021). In contrast, here we use first order vs. higher-order uncertainty in place of risk vs. uncertainty.

Example 1 (COMARG). The COMARG conference is to be held during December 2023. We have the following information concerning the question whether COMARG will be held hybrid (see Figure 1, left).

1. The organizers of COMARG announce that a sufficient condition for the conference to be held hybrid is if there is another wave of COVID in autumn.
2. If there is no COVID wave in autumn, the steering committee will take into account other considerations (such as environmental issue, etc.) and decide on this basis whether the conference is to be held hybrid. We lack any information about how likely it is that such considerations lead to a decision in favor (or disfavor) of a hybrid conference.
3. According to expert opinion, the likelihood of a COVID wave in autumn is at 0.7.

The answer to the question whether the next COMARG conference will be held hybrid is uncertain. Moreover, one cannot attach a precise probability to it: the best that can be said is that it has at least the likelihood 0.7 (given statements 1 and 3). We are dealing with HOU in contradistinction to mere first order uncertainty: in contrast to the question how likely a COVID wave in autumn is, the question how likely it is that COMARG will be held hybrid has no precise answer.

Example 2 (Ellsberg, 1961). Suppose an urn contains 30 red balls, and 60 non-red balls, among which each ball is yellow or black, but we do not know the distribution of yellow and black balls. The question of whether a randomly drawn ball is red is one of first order uncertainty since it comes with the (precise) probability of $1/3$. The question whether it is yellow is one of HOU since the available probabilistic information does not lead to a precise probabilistic estimate. See Figure 1 (right) for an illustration.

1.2. First and higher-order uncertainty in human cognition and AI (HCAI)

Since our environments come with many sources of uncertain information, both quantifiable and not, it is not surprising that human reasoning is well-adjusted to dealing with such situations. What is more, human reasoning distinguishes the two types of uncertainty by treating them differently. For example, in Example 2 people are more willing to bet on drawing a red ball than on drawing a yellow ball in a game in which one wins if one bets the right color. This phenomenon is known as ambiguity (or uncertainty) aversion. The distinction can be traced back both to the psychological and neurological level. For instance, different types of psychological or other medical problems are associated with a compromised decision making under first order uncertainty, but not under HOU [e.g., gambling problems in Brevers et al. (2012), obsessive-compulsive disorder in Zhang et al. (2015), pathological buying issues in Trotzke et al. (2015)] and vice versa (e.g., Parkinson's disease in Euteneuer et al., 2009). This shows that different causal mechanisms are related to the human capacities of reasoning with the two types of uncertainties. Similarly, on the neurological level differences can be traced, though it is still an open issue whether the two uncertainty types have separate or graded

representations in the brain [see De Groot and Thurik (2018), to which we also refer for a recent overview on both the psychological and neurological literature].

What the discussion highlights is that a formal model of human reasoning should pay special attention to both types of uncertainties and provide a framework that can integrate mixed reasoning processes, such as in our Examples 1 and 2. The same can be stated for AI for the simple reason that in many applications artificial agents will face sources of uncertain information.

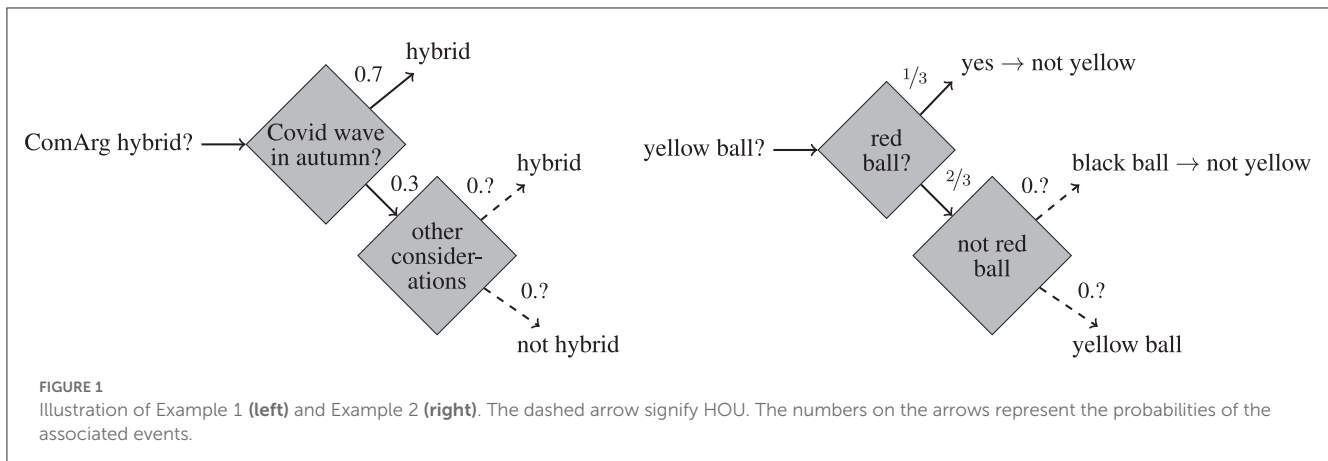
When reasoning with uncertain information, we infer defeasibly, that is, given new (and possibly more reliable) information we may be willing to retract inferences. As forcefully argued on philosophical grounds in Toulmin (1958), reasoning is naturally studied as a form of argumentation. Similarly, the cognitive scientists Mercier and Sperber developed an argumentative theory of human reasoning (Mercier and Sperber, 2017). Dung's abstract argumentation theory (Dung, 1995) provides a unifying formal framework for an argumentative model of defeasible reasoning and has been widely adopted by now both in the context of symbolic AI and to provide explanatory frameworks in the context of human cognition (Saldanha and Kakas, 2019; Cramer and Dietz Saldanha, 2020). Several ways of instantiating abstract argumentation with concrete formal languages and rule sets have been proposed, such as ASPIC+ (Modgil and Prakken, 2014), assumption-based argumentation (Dung et al., 2009), and logic-based argumentation (Besnard and Hunter, 2001; Arieli and Straßer, 2015).

It would therefore seem advantageous for the theoretical foundations of HCAI to combine formal argumentation with a representation of first and higher-order uncertainty. This paper will propose such a formal framework.

1.3. Formal methods

Several formal models of this type of reasoning are available: from imprecise probabilities (Bradley, 2019) to subjective logic (Jøsang, 2001) and probabilistic argumentation (Haenni, 2009). However, the link to the leading paradigm of computational argumentation, namely Dung-style argumentation semantics, is rather loose.

Probabilistic argumentation with uncertain probabilities is comparatively understudied in formal argumentation. Works by Hunter and Thimm (Hunter and Thimm, 2017; Hunter, 2022) focus on precise probabilities. Our framework generalizes aspects of such settings to include a treatment of HOU. Also, in contrast to them, we will utilize Dung argumentation semantics in the context of probabilistic argumentation. Hunter et al. (2020) equip arguments with a degree of belief as well as disbelief, notions that can also be expressed in Haenni's framework and will be considered in our study of argument strength. A framework that considers imprecise probabilities is presented by Oren et al. (2007). It utilizes subjective logic in the context of a dialogical approach for reasoning about evidence. Similarly, Santini et al. (2018) label arguments in abstract argumentation with opinions from subjective logic. In contrast, our study focuses on structured argumentation.



Mainly starting with the seminal (Ellsberg, 1961), HOU has been intensively studied in the context of decision theory. As has been shown there, human reasoning with HOU may lead to violations of axioms of subjective expected utility theory (as axiomatized in Savage, 1972), leading to several alternative accounts [e.g., maxmin expected utility in Gilboa and Schmeidler (2004) or prospect theory in Kahneman and Tversky (1979)]. In this paper, we omit utilities, values, and practical decision making and concentrate instead on reasoning in the epistemic context of belief formation and hypothesis generation. As we will show, even without utilities HOU gives rise to interesting and challenging reasoning scenarios.

1.4. Our contribution

In this paper we integrate reasoning with HOU in abstract argumentation. For achieving this goal, several key questions have to be answered:

1. *What is a knowledge base?* A knowledge base contains strict assumptions (also, constraints) and defeasible assumptions for which probabilistic information is available in form of a family of probability functions. Following Haenni (2009), we distinguish probabilistic and non-probabilistic (also, logical) variables, where only for the former set probabilistic information is available.
2. *What is a logically structured argument?* We will follow the tradition in logical/deductive argumentation according to which an argument is a pair $\langle \mathcal{S}, \phi \rangle$ where \mathcal{S} is a set of assumptions and ϕ a sentence that deductively follows from \mathcal{S} (in signs, $\mathcal{S} \vdash \phi$).
3. *When is an argument stronger than another one?* We propose several measures of argument strength with special consideration of HOU and study their properties. Any model of defeasible reasoning may have various applications, from normative philosophical models of non-monotonic inference to symbolic artificial intelligence, to descriptively adequate (and therefore predictive) models of human reasoning. When considering argument strength, we here focus on the latter and provide a small empirical study (incl. well-known reasoning tasks such as Ellsberg, 1961) to check the accuracy

of the previously defined notions. Clearly, this is a first preliminary step which can only point in a direction, rather than conclusively validate the formal notions developed in this paper.²

4. *What constitutes an argumentative attack?* We study four types of argumentative attack, namely, rebut and three forms of undercut.
5. *How to obtain meta-theoretically well-behaved selections of arguments?* We study several standard argumentation semantics from Dung (1995) for different attack forms in terms of rationality postulates developed for structured argumentation Caminada and Amgoud (2007). When applying argumentation semantics, problems concerning the consistency of extensions already known from logical argumentation re-occur: namely, the set of conclusions of arguments in a given complete extension may be inconsistent. We will propose a solution to this problem that is also applicable in the context of probabilistic argumentation in the style of Hunter and Thimm (2017) and logical argumentation. Moreover, we argue that a naive selection of arguments whose strength passes a certain threshold can lead to inconsistency. Instead, selections in the tradition of Dung seem to be more promising. First, our Dung-based approach satisfies several rationality postulates (including some concerning the consistency of selections). Moreover, it allows for the reinstatement of arguments that are defended by other selected arguments from attacks by non-selected arguments. This is advantageous e.g., when adopting an investigative or hypothetical reasoning style.

Our work takes as the starting point the theory of probabilistic argumentation developed in Haenni (2009). The framework is enhanced by (1) a structured notion of argument in the style of logical argumentation, (2) argumentative attacks, (3) several notions of argument strength [based on notions of degree of support and degree of possibility presented in Haenni (2009)], and (4) a study of Dung-style argumentation

² The focus point of the paper will be on strength measures that associate arguments with strength values in $[0, 1]$, leading to linear strength orderings on the given set of arguments. In future work we will investigate broader notions allowing for incomparabilities between arguments.

semantics. This way, we obtain a generalization of both (some forms of) logical argumentation (Besnard and Hunter, 2018) and probabilistic argumentation in the tradition of Hunter and Thimm (2017).

The structure of the paper is as follows. In Section 2, we introduce knowledge bases and arguments. In Section 3, we discuss the application of argumentation semantics and study rationality postulates relative to the attack form used. Section 4 presents the empirical study on argument strength. We provide a discussion and conclusion in Section 5. In the Appendix (Supplementary material), we provide proofs of our main results, some alternative but equivalent definitions, and details on our empirical study.

2. Knowledge bases and arguments

2.1. Knowledge bases

Our reasoning processes never start from void, but we make use of available information when building arguments. This available information is encoded in a knowledge base. In our initial Example 1, we had two types of information available:

1. probabilistic information concerning a COVID-wave (“the likelihood of a COVID wave in autumn is 70%”). This information may ground defeasible assumption such as “a COVID wave will (probably) (not) take place”; and
2. information about a factual constraint concerning the circumstances in which the conference will be held hybrid (namely, if there is a COVID wave in autumn).

More generally we will follow this rough distinction in probabilistic information that gives rise to defeasible assumptions, on the one hand, and factual constraints, on the other hand. Constraints are taken for granted, either because a reasoner is convinced of their truth, or otherwise committed to them in the reasoning process (e.g., they may be supposed in an episode of hypothetical reasoning³).

Altogether a knowledge base consists of the following components:

Assumptions. Our knowledge bases are equipped with a (finite) set of propositional variables \mathcal{V}_p about which probabilistic information (in the form of probability functions) is available. Out of these propositional variables a set of *defeasible assumptions* \mathcal{A} is formed, whose strength will be influenced by their probabilities. So, where $\text{sent}(\mathcal{V}_p)$ is the set of the propositional sentences with atoms in \mathcal{V}_p , $\mathcal{A} \subseteq \text{sent}(\mathcal{V}_p)$.

Probabilities. We work with a *set probability functions* \mathbb{P} based on $\text{states}(\mathcal{V}_p)$, where $\text{states}(\mathcal{V}_p)$ is the state space for \mathcal{V}_p (i.e., the set of all truth-value assignments $v: \mathcal{V}_p \rightarrow \{0, 1\}$). This allows for cases in which more than one probability function is available, e.g.,

TABLE 1 Overview: technical notation.

Syntactic entities	
p, q, \dots	Propositional atoms
ϕ, ψ, \dots	Propositional sentences
Γ	Set of sentences
\hat{s}	Syntactic representation of state s
\mathcal{V}_p	Set of probabilistic propositional variables
\mathcal{V}_l	Set of logical propositional variables
$\text{sent}(\mathcal{V})$	Set of sentences over \mathcal{V}
\mathcal{A}	Set of assumptions (subset of $\text{sent}(\mathcal{V}_p)$)
\mathcal{C}	Set of constraints (subset of $\text{sent}(\mathcal{V}_p \cup \mathcal{V}_l)$)
\mathbb{K}	Knowledge base
a, b, \dots	Arguments
$\text{Sup}(a)$	Support of a
$\text{Con}(a)$	Conclusion of a
$\text{Arg}(\mathbb{K})$	Set of arguments induced by \mathbb{K}
$@(\Gamma)$	The argument $(\Gamma, \bigwedge \Gamma)$
\mathcal{E}	Set of arguments, an argumentation extension
Semantic entities	
s	State
$\text{states}(\mathcal{V})$	Set of states over variables in \mathcal{V}
P	Probability function
$P^{\mathcal{C}}$	Probability function after Bayesian update on \mathcal{C}
\mathbb{P}	Set of probability functions
$\wp(\cdot)$	Power set
$\ \phi\ $	Set of states that verify ϕ
$\ \phi\ _{\mathcal{C}}$	Set of states that verify ϕ and are consistent with \mathcal{C}
$\text{str}(a)$	Placeholder for argument strength function
$\text{dsp}(a)$	Degree of support of a
$\text{dps}(a)$	Degree of possibility of a

scenarios in which multiple sources of probabilities are considered or in which the probabilistic information about the state space $\text{states}(\mathcal{V}_p)$ is incomplete (see below for examples).

For a sentence ϕ we let $\|\phi\|$ be the set of states $s \in \text{states}(\mathcal{V}_p)$ for which $s \models \phi$. For some $s \in \text{states}(\mathcal{V}_p)$, we denote by \hat{s} the conjunction $\bigwedge \{\phi \mid \alpha(\phi) = 1, \phi \in \mathcal{V}_p\} \cup \{\neg\phi \mid \alpha(\phi) = 0, \phi \in \mathcal{V}_p\}$. The reader finds an overview of the notation used in this paper in Table 1.

We may *also* use propositional variables for which no (direct) probabilistic information is considered. We collect these in \mathcal{V}_l (the logical variables) and require $\mathcal{V}_l \cap \mathcal{V}_p = \emptyset$. By allowing logical variables as well as probabilistic ones, we can unite logical (where $\mathcal{V}_p = \emptyset$) and probabilistic reasoning (where $\mathcal{V}_l = \emptyset$) and can involve both systems seamlessly, following the approach by Haenni

³ For work in non-probabilistic structured argumentation on hypothetical reasoning, see Beirlaen et al. (2018), Cramer and Dauphin (2019).

TABLE 2 The state space and the probabilities for Example 3.

State	$s \models p$	P
s_1		0.3
s_2	✓	0.7

(2009). Constraints will typically relate probabilistic with non-probabilistic information and therefore they are based on atoms from both sets, \mathcal{V}_p and \mathcal{V}_l .

Constraints. The last element of knowledge bases is a set of factual constraints (in short, *constraints*) \mathcal{C} . The formulas in \mathcal{C} are based on atoms in $\mathcal{V}_p \cup \mathcal{V}_l$.

In the following we write $\Gamma \vdash_{\mathcal{C}} \phi$ as an abbreviation of $\Gamma \cup \mathcal{C} \vdash \phi$, where $\Gamma \subseteq \text{sent}(\mathcal{V}_l \cup \mathcal{V}_p)$. It will also be useful to collect all states that are consistent with \mathcal{C} and that support a formula $\phi \in \text{sent}(\mathcal{V}_l \cup \mathcal{V}_p)$ in $\|\phi\|_{\mathcal{C}} = \{s \in \text{states}(\mathcal{V}_p) \mid s \not\models_{\mathcal{C}} \neg\phi \text{ and } s \vdash_{\mathcal{C}} \phi\}$. Similarly, we write $s \models_{\mathcal{C}} \phi$ in case $s \in \|\phi\|_{\mathcal{C}}$.

We summarize the above discussion in the following definition:

Definition 1 (Knowledge Base). A knowledge base \mathbb{K} is a tuple $\langle \langle \mathcal{V}_p, \mathcal{V}_l \rangle, \mathcal{A}, \mathcal{C}, \mathbb{P} \rangle$ for which

- \mathcal{V}_p is a finite set of *probabilistic variables*,
- \mathcal{V}_l is a finite set of *logical variables* such that $\mathcal{V}_l \cap \mathcal{V}_p = \emptyset$,⁴
- $\mathcal{A} \subseteq \text{sent}(\mathcal{V}_p)$ is a finite set of *defeasible assumptions*,
- $\mathcal{C} \subseteq \text{sent}(\mathcal{V}_p \cup \mathcal{V}_l)$ is a finite set of *constraints*,
- \mathbb{P} is a non-empty set of probability functions $P : \text{states}(\mathcal{V}_p) \rightarrow [0, 1]$.

Example 3 (Ex. 1 cont.). Let us return to our example. It can be modeled by the knowledge base

$$\text{COVID} = \langle \langle \mathcal{V}_p : \{p\}, \mathcal{V}_l : \{q\} \rangle, \mathcal{A} : \{p, \neg p\}, \mathcal{C} : \{p \rightarrow q\}, \mathbb{P} : \{P\} \rangle,$$

where $P(p) = 0.7$, p stands for a COVID-Wave to happen and q for the conference to be held hybrid. Our defeasible assumptions \mathcal{A} are $\{p, \neg p\}$ (“there will (not) be a COVID-Wave”). Our set of factual constraints is $\mathcal{C} = \{p \rightarrow q\}$. Table 2 shows the state space induced by \mathcal{V}_p .

2.2. Arguments, support, and strength

Given a knowledge base $\mathbb{K} = \langle \mathcal{A}, \mathcal{C}, \mathbb{P} \rangle$, a natural way of thinking about arguments is in terms of support-conclusion pairs:

Definition 2 (Argument). Given a knowledge base $\mathbb{K} = \langle \mathcal{A}, \mathcal{C}, \mathbb{P} \rangle$, an *argument* a for \mathbb{K} is a pair $\langle \text{Sup}(a), \text{Con}(a) \rangle$, where

- $\text{Sup}(a) \subseteq \mathcal{A}$ is a set of assumptions, the *support* of a ,
- $\text{Con}(a) \in \text{sent}(\mathcal{V}_l \cup \mathcal{V}_p)$ is the *conclusion* of a ,

⁴ Where the context disambiguates, we will omit the listing of the variables \mathcal{V}_p and \mathcal{V}_l to avoid clutter in the notation.

- such that $\text{Sup}(a) \vdash_{\mathcal{C}} \text{Con}(a)$.

We write $\text{Arg}(\mathbb{K})$ for the set of all arguments based on \mathbb{K} .

Example 4 (Ex. 3 cont.). In our example we can form the argument $\text{hybrid} = \langle \{p\}, q \rangle$ for the conference to be held hybrid, the argument $\text{wave} = \langle \{p\}, p \rangle$ for there being a COVID-wave, and $\text{noWave} = \langle \{\neg p\}, \neg p \rangle$ for there being no wave.

When considering the question of how strong an argument $a = \langle \Gamma, \phi \rangle$ is, a naive approach is to simply measure the probabilistic strength of the support. In the simple case of our example and the argument $\text{hybrid} = \langle \{p\}, q \rangle$ this would amount to $P(p) = 0.7$, the same as for the argument wave , whereas noWave would only have a strength of 0.3. However, there are some subtleties which motivate a more fine-grained analysis. To show this, we enhance our example as follows.

Example 5 (COMARG2). We also consider another conference, COMARG2, for which we know that it will be held hybrid (symbolized by q') if and only if(!) a COVID-wave breaks in autumn. Our enhanced knowledge base is $\text{COVID}' = \langle \langle \{p\}, \{q, q'\} \rangle, \mathcal{A} : \{p, \neg p\}, \mathcal{C}' : \{p \rightarrow q, p \leftrightarrow q'\} \rangle$. We now added also $p \leftrightarrow q'$ to the set of constraints \mathcal{C} . We can now also consider the additional argument $\text{hybrid}' = \langle \{p\}, q' \rangle$ for COMARG2 to be held hybrid.

Observation 1 (Stronger, but less precise arguments.). Intuitively, the argument hybrid in favor of q is stronger than the argument hybrid' in favor of q' (see also our empirical study in Section 4). Although both arguments have the same support, $\{p\}$, the conclusion q of hybrid is compatible with both states, s_1 and s_2 (it is certain in s_2 and possible in s_1), while the conclusion q' of hybrid' is only compatible with s_2 . As a consequence, q has *at least* the probability 0.7, while q' has *precisely* the probability 0.7.

In sum, it is intuitive to consider an argument a as at least as strong as an argument b , in case both arguments have the same support, but the conclusion of a is at least as probable as the conclusion of b .

Let us analyse this observation in more formal terms. We write $s \models_{\mathcal{C}} \Diamond\phi$ iff $s \in \|\top\|_{\mathcal{C}} \setminus \|\neg\phi\|_{\mathcal{C}}$. This means that ϕ is possible in s in view of the constraints in \mathcal{C} . Similarly, we write $\|\Diamond\phi\|_{\mathcal{C}}$ for the set of states $\|\top\|_{\mathcal{C}} \setminus \|\neg\phi\|_{\mathcal{C}}$.

Fact 1. Let $a \in \text{Arg}(\mathbb{K})$ and $\text{Con}(a) \vdash_{\mathcal{C}} \phi$. Then (1) $\|\text{Sup}(a)\|_{\mathcal{C}} = \|\bigwedge \text{Sup}(a)\|_{\mathcal{C}} \subseteq \|\text{Con}(a)\|_{\mathcal{C}} \subseteq \|\phi\|_{\mathcal{C}}$, and (2) $\|\text{Con}(a)\|_{\mathcal{C}} \subseteq \|\Diamond\text{Con}(a)\|_{\mathcal{C}} \subseteq \|\Diamond\phi\|_{\mathcal{C}}$.

In our Example 5 we have the validities for the different states shown in Table 3. Following Observation 1, $\text{hybrid} = \langle \{p\}, q \rangle$ is stronger than $\text{hybrid}' = \langle \{p\}, q' \rangle$. The reason seems to be that despite having the same support, the “space of possibility” for q is larger than the one for q' : $\{s_1, s_2\}$ vs. $\{s_2\}$. From the probabilistic perspective, the support for p seems to be located in $[0.7, 1]$ while the one for q' is exactly 0.7.

Following this rationale, the strength of the support of an argument is measured relative to a lower and upper bound: the

TABLE 3 Validities for Example 5.

State s	$s \models p$	$s \models_{\mathcal{C}'} q$	$s \models_{\mathcal{C}'} \Diamond q$	$s \models_{\mathcal{C}'} q'$	$s \models_{\mathcal{C}'} \Diamond q'$
s_1			✓		
s_2	✓	✓	✓	✓	✓

lower bound is the cautious measure of how probable the support is in the *worst case*, the upper bound considers the *best case* scenario in which states in which the conclusion holds have maximal probability mass. As we will see, the central idea for modeling argument strength in this paper is by means of functions that map arguments to $[0, 1]$ (their strength) by aggregating the worst and best case support.

Before discussing two complications, we shortly summarize the ideas so far. Arguments are support-conclusion pairs. When considering the strength of an argument $a = \langle \text{Sup}, \phi \rangle$ it is advisable not only to consider the probabilistic strength of its support Sup , but also to consider the probabilistic support for the possibility of its conclusion $\Diamond \phi$. A measure of argument strength is expected to aggregate the two.

2.3. Imprecise probabilistic information

In many scenarios it will be advantageous or unavoidable to work with families of probability functions, instead of a unique probability function. These are cases in which the probabilistic information concerning the probabilistic variables in \mathcal{V}_p is incomplete or it stems from various sources, each providing an individual probability function. The following example falls in the former category.

Example 6 (Three conferences). Peter and Mary are in the steering committee of CONFB, CONFP, and CONFM. Their votes have different weights for the decision making of the respective committees. Both of their positive votes are sufficient but not necessary for CONFB to be held hybrid. For CONFP the decision relies entirely on Peter's vote, and for CONFM it relies entirely on Mary's vote.

- If Peter votes hybrid, CONFB will be hybrid. $p_1 \rightarrow q_1$
 - If Mary votes hybrid, CONFB will be hybrid. $p_2 \rightarrow q_1$
 - CONFP will be hybrid if and only if Peter votes hybrid. $p_1 \leftrightarrow q_2$.
 - CONFM will be hybrid if and only if Mary votes hybrid. $p_2 \leftrightarrow q_3$.
 - According to Peter, there is a $2/3$ likelihood that he will vote hybrid. $P(p_1) = 2/3$
 - According to Mary, she is at least as likely to vote hybrid as Peter. $P(p_2) \geq 2/3$
- (We lack more precise information.)

Altogether our knowledge base is given by $\mathbb{K} = \langle \{p_1, p_2\}, \{q_1, q_2, q_3\}, \mathcal{A}, \mathcal{C}, \mathbb{P} : \{P_\mu \mid \mu \in [0, 1/3]\} \rangle$, where $\mathcal{A} = \text{sent}(\mathcal{V}_p)$ and $\mathcal{C} = \{p_1 \rightarrow q_1, p_2 \rightarrow q_1, p_1 \leftrightarrow q_2, p_2 \leftrightarrow q_3\}$.

Moreover, in this case the probabilities for our defeasible assumptions p_1 and p_2 are not precise. They are expressed by means of a family of probability functions (see Table 4).⁵

Given an argument $\langle \text{Sup}, \phi \rangle$, a cautious way to consider the worst case probabilistic support is by considering $\inf_{P \in \mathbb{P}} (P(\|\text{Sup}\|_{\mathcal{C}}))$. Following Haenni, we refer to this measure as the *degree of support* of an argument. For the best case probabilistic support, on the other hand, we consider $\sup_{P \in \mathbb{P}} (P(\|\Diamond \phi\|_{\mathcal{C}}))$. We refer to this measure as the *degree of possibility* of an argument. An overview for the current example can be found in Table 5. Before formally defining the two discussed measures, we have to still consider one more complication, however, which will discuss in the next section.

2.4. Updating the probabilities in view of the constraints

Consider the following example:

- Example 7** (Witnesses). 1. According to witness 1 $p \wedge q$ is the case. $p_1 \rightarrow p \wedge q$
2. According to witness 2 $p \wedge \neg q$ is the case. $p_2 \rightarrow p \wedge \neg q$
3. Witness 1 tells the truth in $2/3$ of cases. $P(p_1) = 2/3$
4. Witness 2 tells the truth in $3/4$ of cases. $P(p_2) = 3/4$

We may model this scenario with the knowledge base $\mathbb{K} = \langle \langle \mathcal{V}_p : \{p_1, p_2\}, \mathcal{V}_l : \{p, q\} \rangle, \mathcal{A} : \text{sent}(\mathcal{V}_p), \mathcal{C} : \{p_1 \rightarrow p \wedge q, p_2 \rightarrow p \wedge \neg q\}, \mathbb{P} = \{P\} \rangle$ where P assigns the probabilities as depicted in Table 6.

In this case s_4 is incompatible with the set of constraints \mathcal{C} of our knowledge in \mathbb{K} and the probabilities have to be updated. We follow Haenni (2009) by using a Bayesian update on $\bigwedge \mathcal{C}$ and letting

$$P^{\mathcal{C}}(s) = \frac{P(s)}{P(\|\mathcal{C}\|)} \cdot P(\|\mathcal{C}\| \mid s) = \frac{P(s)}{P(\|\top\|_{\mathcal{C}})} \cdot \frac{P(\|s\|_{\mathcal{C}})}{P(s)} = \frac{P(\|s\|_{\mathcal{C}})}{P(\|\top\|_{\mathcal{C}})}. \quad (1)$$

Similarly, where \mathbb{P} is a family of probability functions, we let $\mathbb{P}^{\mathcal{C}} = \{P^{\mathcal{C}} \mid P \in \mathbb{P}\}$. When calculating the degrees of support and degrees of possibility of an argument we will consider $\mathbb{P}^{\mathcal{C}}$ instead of \mathbb{P} .

Definition 3 (Degree of Support and Degree of Possibility, (Im)Precision). Given a knowledge base $\mathbb{K} = \langle \mathcal{A}, \mathcal{C}, \mathbb{P} \rangle$ and an argument $a = \langle \text{Sup}, \phi \rangle$ for \mathbb{K} ,

- The degree of support of a (in signs, $\text{dsp}(a)$) is given by $\inf_{P \in \mathbb{P}^{\mathcal{C}}} (P^{\mathcal{C}}(\|\text{Sup}\|))$,
- The degree of possibility of a (in signs, $\text{dps}(a)$) is given by $\sup_{P \in \mathbb{P}^{\mathcal{C}}} (P^{\mathcal{C}}(\|\Diamond \phi\|))$,
- The imprecision of a (in signs, $\text{imprec}(a)$) is given by $\text{dps}(a) - \text{dsp}(a)$,
- The precision of a (in signs, $\text{prec}(a)$) is given by $1 - \text{imprec}(a)$.

Fact 2. Let $a, b \in \text{Arg}(\mathbb{K})$.

⁵ Here we assume probabilistic independence of p_1 and p_2 . If this assumption is given up we operate on the basis of $\mathbb{P} = \{P_{\lambda, \mu} \mid \lambda \in [0, 2/3], \mu \in [2/3 - \lambda, 1/3]\}$ where $P_{\lambda, \mu} : s_1 \mapsto 1/3 - \mu, s_2 \mapsto \mu, s_3 \mapsto 2/3 - \lambda, s_4 \mapsto \lambda$.

TABLE 4 The state space and probabilities for Example 6, where $\mu \in [0, 1/3]$.

State	p_1	p_2	P_μ	$P_{\mu=0}$	$P_{\mu=1/3}$	q_1	q_2	q_3
s_1	0	0	$1/3 \cdot (1/3 - \mu)$	$1/9$	0	\diamond	0	0
s_2	0	1	$1/3 \cdot (2/3 + \mu)$	$2/9$	$1/3$	1	0	1
s_3	1	0	$2/3 \cdot (1/3 - \mu)$	$2/9$	0	1	1	0
s_4	1	1	$2/3 \cdot (2/3 + \mu)$	$4/9$	$2/3$	1	1	1

TABLE 5 The degrees of support and possibility for Example 6.

Argument	Degree of support	Degree of possibility	Precision	Imprecision
$a_1 = \langle p_1, q_1 \rangle$	$\inf_{\mu \in [0, 1/3]} (P_\mu(\ p_1\)) = \inf(\{2/3\}) = 2/3$	$\sup_{\mu \in [0, 1/3]} (P_\mu(\ \diamond q_1\)) = \sup(\{1\}) = 1$	$2/3$	$1/3$
$a_2 = \langle p_2, q_1 \rangle$	$\inf_{\mu \in [0, 1/3]} (P_\mu(\ p_2\)) = \inf(\{2/3, 1\}) = 2/3$	$\sup_{\mu \in [0, 1/3]} (P_\mu(\ \diamond q_1\)) = \sup(\{1\}) = 1$	$2/3$	$1/3$
$a_3 = \langle p_1 \vee p_2, q_1 \rangle$	$\inf_{\mu \in [0, 1/3]} (P_\mu(\ p_1 \vee p_2\)) = \inf(\{8/9, 1\}) = 8/9$	$\sup_{\mu \in [0, 1/3]} (P_\mu(\ \diamond q_1\)) = \sup(\{1\}) = 1$	$8/9$	$1/9$
$b = \langle p_1, q_2 \rangle$	$\inf_{\mu \in [0, 1/3]} (P_\mu(\ p_1\)) = \inf(\{2/3\}) = 2/3$	$\sup_{\mu \in [0, 1/3]} (P_\mu(\ \diamond q_2\)) = \sup(\{2/3\}) = 2/3$	1	0
$c = \langle p_2, q_3 \rangle$	$\inf_{\mu \in [0, 1/3]} (P_\mu(\ p_2\)) = \inf(\{2/3, 1\}) = 2/3$	$\sup_{\mu \in [0, 1/3]} (P_\mu(\ \diamond q_3\)) = \sup(\{2/3, 1\}) = 1$	$2/3$	$1/3$

TABLE 6 The states for Example 7.

State s	p_1	p_2	P	$s \in \ \top\ _{\mathcal{C}}$	$P^{\mathcal{C}}(s)$	$s \models_{\mathcal{C}} p$	$s \models_{\mathcal{C}} \diamond p$
s_1	0	0	$1/12$	✓	$P(s_1)/P(\{s_1, s_2, s_3\}) = 1/6$		✓
s_2	0	1	$1/4$	✓	$P(s_2)/P(\{s_1, s_2, s_3\}) = 1/2$	✓	✓
s_3	1	0	$1/6$	✓	$P(s_3)/P(\{s_1, s_2, s_3\}) = 1/3$	✓	✓
s_4	1	1	$1/2$		$P(s_4)/P(\{s_1, s_2, s_3\}) = 0$		

The 5th column indicates which states are consistent with \mathcal{C} (the only exception is state s_4). The 6th column represents the updated probabilities for each state in accordance with Equation (1).

1. If $\text{Sup}(a) \subseteq \text{Sup}(b)$ then $\text{dsp}(a) \geq \text{dsp}(b)$.
2. If $\{\text{Con}(a)\} \vdash_{\mathcal{C}} \text{Con}(b)$ then $\text{dps}(b) \geq \text{dps}(a)$.

As discussed above, we expect a measure of argument strength to aggregate the two measures of degree of support and degree of possibility.

Definition 4 (Argument strength function). Let $\mathbb{K} = \langle \mathcal{A}, \mathcal{C}, \mathbb{P} \rangle$ be a knowledge base. A measure of argument strength for \mathbb{K} is a function $\text{str} : \text{Args}(\mathbb{K}) \rightarrow [0, 1]$ that is associated with a function $\pi : \Theta \rightarrow [0, 1]$ for which $\Theta = \{(n, m) \in [0, 1]^2 \mid n \leq m\}$ and $\text{str}(a) = \pi(\text{dsp}(a), \text{dps}(a))$.

3. Argument selection

In this section, we consider the question of how to evaluate the strength of arguments and how to select them for acceptance out of a scenario of possibly conflicting arguments. The questions of argument strength and of argument selection are connected: e.g., if two arguments conflict, it is usually advisable to select the stronger of the two. We will proceed in several steps.

1. We propose several notions of argument strength and study their properties (Section 3.1).
2. In Section 3.2, we discuss two types of argumentative attacks: rebuttals and undercuts. We show that both lead to suboptimal outcomes when combined with Dung-style argumentation semantics for selecting arguments in a naive way.
3. In Section 3.3, we propose a solution to the problem of argument selection.

While this section is devoted to the theoretic foundations of probabilistic argumentation, we will provide a small empirical study to compare some of the proposed measures in Section 4.

3.1. Argument strength

As discussed above, we have two underlying measures which can serve as input for a measure of argument strength: the degree of support and the degree of possibility (recall Definition 4): $\text{str}(a) = \pi(\text{dsp}(a), \text{dps}(a))$ where $\pi : \{(n, m) \in [0, 1]^2 \mid n \leq m\} \rightarrow [0, 1]$. As for π there are various straight-forward options. We list a few in Table 7. *Support* and *possibility* reflect the lower and upper

probabilistic bounds represented by dsp and dps , while *mean* represents their mean. *Boosted support* follows the idea underlying Observation 1 according to which an argument c with $\text{dsp}(c) < \text{dps}(c)$ should get a “boost” as compared to an argument d for which $\text{dsp}(d) = \text{dps}(d) = \text{dsp}(c)$. The factor $m \geq 1$ determines the magnitude of the boost, the lower m the more the lower bound dsp is boosted (where for $m = 1$ the boosted support is identical to dps). *Convex combination* follows a similar idea by letting the strength of an argument a be the result of a convex combination of $\text{dsp}(a)$ and $\text{dps}(a)$, where the parameter α determines how cautious an agent is: the higher α the less epistemic risk an agent is willing to take (where for $\alpha = 1$ the convex combination is identical to $\text{dsp}(a)$). *Precision mean* is a qualification of mean in that it also considers the precision of an argument as a marker of strength (see Pfeifer, 2013). The precision of an argument a is given by $1 - (\text{dps}(a) - \text{dsp}(a))$: the closer $\text{dsp}(a)$ and $\text{dps}(a)$ the more precise is a . The *precision mean* of an argument is the result of multiplying its mean with its precision. We note that this measure is in tension with the intuition behind Observation 1 in that it would measure the strength of *hybrid* higher than that of *hybrid'*, unlike *boosted support* or *convex combination*.

Clearly, some of the measures coincide for specific parameters (the proof can be found in Appendix A):

- Fact 3.** 1. $\text{mean}(a) = \text{bst}_2(a) = \text{convex}_{.5}(a)$
 2. $\text{dsp}(a) = \text{convex}_1(a)$ and $\text{dps}(a) = \text{convex}_0(a) = \text{bst}_1(a)$
 3. $\text{bst}_m(a) = \text{convex}_{1-1/m}(a)$ and $\text{convex}_\alpha(a) = \text{bst}_{1/(1-\alpha)}(a)$ (where $\alpha < 1$).

Proof: Items 1 and 2 are trivial. We show Item 3. We have, on the one hand, $\text{bst}_m(a) = \text{dsp}(a) + \frac{\text{dps}(a) - \text{dsp}(a)}{m} = \text{dsp}(a) - \frac{\text{dsp}(a)}{m} + \frac{\text{dps}(a)}{m} = (1 - 1/m) \cdot \text{dsp}(a) + (1 - (1 - 1/m)) \cdot \text{dps}(a) = \text{convex}_{1-1/m}(a)$. On the other hand, $\text{convex}_\alpha(a) = \alpha \cdot \text{dsp}(a) + (1 - \alpha) \text{dps}(a) = \text{dsp}(a) + \text{dps}(a) - \text{dsp}(a) \cdot \alpha - \text{dsp}(a) + \text{dsp}(a) \cdot \alpha = \text{dsp}(a) + (\text{dps}(a) - \text{dsp}(a)) \cdot (1 - \alpha) = \text{bst}_{1/(1-\alpha)}(a)$. \square

Example 8. In Table 8, we apply the different argument strength measures to Examples 1 and 6.

We now analyse the different strength measures in view of several properties, some of which may be considered desirable.⁶ Table 9 offers an overview on which properties are satisfied for which measures.

- **Domain Restriction.** $\text{str}(a) \in [\text{dsp}(a), \text{dps}(a)]$. In the context of a given knowledge base, the degree of support represents a cautious estimation of the probability of the conclusion of a in view of its support, while the degree of possibility represents the most optimistic (in that it considers its possibility) estimation of its probability.
- **Precision.** If $\text{prec}(a) = 1$ then $\text{str}(a) = \text{dsp}(a) = \text{dps}(a)$. This is a special case of Domain Restriction for cases in which the available information concerning a is precise.

- **Neutrality.** $\text{str}(a) = 0.5$ if $\text{prec}(a) = 0$. If $\text{prec}(a) = 0$, we have $\text{dsp}(a) = 0$ and $\text{dps}(a) = 1$. According to Neutrality we treat such cases as flipping an unbiased coin.
- **Moderation.** $\text{str}(a) \leq \text{mean}(a)$. Moderation is a cautious approach, putting more weight on the degree of support than the degree of possibility.

The following properties specify various ways the degrees of support and/or possibility are related to argument strength in terms of offering sufficient resp. necessary conditions. For the following properties let $a \sqsubseteq b$ iff $\text{dsp}(a) \leq \text{dsp}(b)$ and $\text{dps}(a) \leq \text{dps}(b)$. Let \sqsubset be the strict version of \sqsubseteq , i.e., $a \sqsubset b$ iff $a \sqsubseteq b$ and $b \not\sqsubseteq a$.

Fact 4. Let a, b be precise arguments (so, $\text{prec}(a) = \text{prec}(b) = 1$). If Precision holds for str , then: $\text{str}(a) \leq \text{str}(b)$ iff $a \sqsubseteq b$.

- **Weak epistemic sufficiency.** $\text{str}(a) \leq \text{str}(b)$ if $a \sqsubseteq b$.
- **Strict epistemic sufficiency.** $\text{str}(a) < \text{str}(b)$ if $a \sqsubset b$. Our Observation 1 follows the intuition of Strict epistemic sufficiency. In Example 5 we have $\text{hybrid} \sqsubset \text{hybrid}'$ and therefore we expect also $\text{str}(\text{hybrid}) > \text{str}(\text{hybrid}')$.
- **Epistemic risk aversion.** $\text{dsp}(a) \leq \text{dsp}(b)$ if $\text{str}(a) \leq \text{str}(b)$. The criterion says that for b to be at least as strong as a it also has to have an at least as strong degree of support. The agent would take epistemic risk if it were to consider an argument b stronger than a , although b has less degree of support (but maybe more degree of possibility). The contrast case is expressed next.⁷
- **Epistemic risk tolerance.** It is possible that $\text{str}(a) \leq \text{str}(b)$ while $\text{dsp}(a) > \text{dsp}(b)$.
- **Upper compensation.** $\text{str}(a) > \text{str}(b)$ and $\text{mean}(a) \leq \text{mean}(b)$ implies $\text{dps}(a) > \text{dps}(b)$. Choosing an argument a over b despite the fact that b has at least as high mean has to be compensated by a having a higher degree of possibility.
- **Lower compensation.** $\text{str}(a) > \text{str}(b)$ and $\text{mean}(a) \leq \text{mean}(b)$ implies $\text{dsp}(a) > \text{dsp}(b)$. Analogous to the previous criterion, except that the compensation is in terms of the degree of support.

The following criteria present various ways of considering precision a sign of argument quality. For instance, Pfeifer (2013) considers precision a central marker of strength.

- **Precision sufficiency.** If $\text{mean}(a) = \text{mean}(b)$ and $\text{prec}(a) \geq \text{prec}(b)$ then $\text{str}(a) \geq \text{str}(b)$. If two arguments have the same mean, the one with more precision is better. The rationale is that the latter is supported by more informative evidence.
- **Strict precision sufficiency.** If $\text{mean}(a) = \text{mean}(b)$ and $\text{prec}(a) > \text{prec}(b)$ then $\text{str}(a) > \text{str}(b)$.
- **Precision necessity.** $\text{str}(a) \geq \text{str}(b)$ implies $\text{prec}(a) \geq \text{prec}(b)$. An argument can only be at least as good as another one if its precision is at least as good.

⁶ The question of what properties are considered desired depends on the applications: if the application is to obtain a predictive mode of human reasoning these properties are in need of empirical verification (see Section 4).

⁷ We note that Epistemic risk aversion is not very suitable for strength measures that linearly order arguments such as the ones studied in this paper. Indeed, dsp is the only of our measures that satisfies it.

TABLE 7 Various notions of argument strength expressed as function of the degree of support and the degree of possibility of an argument.

Name	$\pi : (x, y) \mapsto \dots$	$\text{str}(a) = \pi(\text{dsp}(a), \text{dps}(a)) = \dots$
Support	x	$\text{dsp}(a)$
Possibility	y	$\text{dps}(a)$
Mean	$\frac{x+y}{2}$	$\text{mean}(a) = \frac{\text{dsp}(a) + \text{dps}(a)}{2}$
Boosted support	$x + \frac{y-x}{m} \ (m \geq 1)$	$\text{bst}_m(a) = \text{dsp}(a) + \frac{\text{imprec}(a)}{m}$
Convex combination	$\alpha \cdot x + (1 - \alpha) \cdot y \ (\alpha \in [0, 1])$	$\text{convex}_\alpha(a) = \alpha \cdot \text{dsp}(a) + (1 - \alpha) \cdot \text{dps}(a)$
Precision mean	$\frac{x+y}{2} \cdot (1 - (y - x))$	$\text{precMean}(a) = \text{mean}(a) \cdot \text{prec}(a)$

TABLE 8 The strengths of arguments presented in Examples 5 and 6.

Example	Argument	$\text{dsp}(a)$	$\text{dps}(a)$	$\text{mean}(a)$	$\text{bst}_3(a)$ resp. $\text{convex}_{2/3}(a)$	$\text{precMean}(a)$
Example 5	wave	0.7	0.7	0.7	0.7	0.7
	hybrid	0.7	1	0.85	0.8	0.595
	hybrid'	0.7	0.7	0.7	0.7	0.7
Example 6	a_1	$2/3$	1	$5/6$	$7/9$	$5/9$
	a_2	$2/3$	1	$5/6$	$7/9$	$5/9$
	a_3	$8/9$	1	$17/18$	$25/27$	$68/81$
	b	$2/3$	$2/3$	$2/3$	$2/3$	$2/3$
	c	$2/3$	1	$5/6$	$7/9$	$5/9$

TABLE 9 Overview on the properties.

Property	$\text{dsp}(a)$	$\text{dps}(a)$	$\text{mean}(a)$	$\text{bst}_m(a)$	$\text{convex}_\alpha(a)$	$\text{precMean}(a)$
Domain restriction [†]	✓	✓	✓	✓	✓	✗ [Example 10]
Precision [†]	✓	✓	✓	✓	✓	✓
Neutrality [†]	✗ [Example 9]	✗ [Example 9]	✓	✓ [$m = 2$]	✓ [$\alpha = 0.5$]	✗
Moderation [†]	✓	✗ [Example 9]	✓	✓ [$m \geq 2$]	✓ [$\alpha \geq 0.5$]	✓
Weak ep. sufficiency [♡]	✓	✓	✓	✓	✓	✗ [Example 10]
Strict ep. sufficiency	✗ [Example 9]	✗ [Example 9]	✓ [†]	✗ [Example 9] [◊]	✗ [Example 9] [◊]	✗ [Example 10]
Ep. risk aversion	✓	✗ [Example 12]	✗ [Example 12]	✗ [Example 12]	✗ [Example 12]	✗ [Example 12]
Ep. risk tolerance	✗	✓ [Example 12]	✓ [Example 12]	✓ [Example 12]	✓ [$\alpha < 1$]	✓
Upper compensation [•]	✗ [Example 9]	✓	✓ [†]	✓	✓ [$\alpha < 1$]	✗ [Example 10]
Lower compensation [•]	✓	✗ [Example 9]	✓	✓ [$m \geq 2$]	✓ [$\alpha \geq 0.5$]	✗ [Example 10]
Precision sufficiency [‡]	✓	✗ [Example 9]	✓	✓ [$m \geq 2$]	✓ [$\alpha \leq 0.5$]	✓
Str. prec. sufficiency [‡]	✓	✗ [Example 9]	✗ [Example 9]	✓ [$m > 2$]	✓ [$\alpha < 0.5$]	✓
Precision necessity	✗ [Example 9]	✗ [Example 9]	✗ [Example 9]	✗ [Example 9] [◊]	✗ [Example 9] [◊]	✗ [Example 10]
Precision compensation	✓ [Proposition 4]	✗ [Example 9]	✓ [†]	✗ [Example 9] [◊]	✗ [Example 9] [◊]	✓ [Proposition 4]
Counter [‡]	✓	✓	✓	✓	✓	✓
R-Weakening [*]	✓	✓	✓	✓	✓	✗ [Example 10]
L-Weakening [*]	✓	✓	✓	✓	✓	✓

(†) The proofs of these properties are trivial and therefore omitted. (♡) shown in Proposition 2. (•) shown in Proposition 6. (‡) shown in Proposition 3. (§) shown in Proposition 7. (★) shown in Proposition 8. (●) shown in Proposition 5. (◊) The counter-examples for dsp and mean apply in view of Fact 3. Proposition 2–8 and their proofs are presented in [Appendix A \(Supplementary material\)](#).

- **Precision compensation.** $\text{str}(a) > \text{str}(b)$ and $\text{mean}(a) \leq \text{mean}(b)$ implies $\text{prec}(a) > \text{prec}(b)$. Choosing an argument a over b despite the fact that b has at least as high mean, has to be compensated by a having a higher precision.

Finally, we offer some criteria that relate arguments to other arguments in a logical way.

- **Counter.** If $\inf_{p \in \mathbb{P}} (P(\|\text{Con}(a)\|_C)) = 0$ and $\text{Con}(b) = \neg \text{Con}(a)$, then $\text{str}(b) \geq \text{str}(a)$. If the conclusion of a has no probabilistic support in the knowledge base and b concludes the opposite, then b is at least as good as a .
- **R-Weakening.** If $\text{Sup}(a) = \text{Sup}(b)$ and $\text{Con}(a) \vdash_C \text{Con}(b)$ then $\text{str}(b) \geq \text{str}(a)$. For two arguments with the same support the one with the logically weaker conclusion is at least as strong as the other argument. Clearly, its conclusion is more cautious.
- **L-Weakening.** If $\text{Sup}(a) \supseteq \text{Sup}(b)$ and $\text{Con}(a) = \text{Con}(b)$ then $\text{str}(a) \leq \text{str}(b)$. For two arguments with the same conclusions the argument which has more support is at most as strong as the other argument.

Before studying these properties for our different notions of argument strength, we observe some logical relations between some of them.

Proposition 1. For any argument strength measure str we have:

1. If str satisfies Domain restriction then it satisfies Precision.
2. If str satisfies Weak epistemic sufficiency, then it also satisfies R-weakening and L-weakening.

Proof: Ad 1. Trivial. Ad 2. Concerning R-weakening and L-weakening, observe that if a and b fulfill the requirements of the left hand side of R-weakening resp. of L-weakening, then $b \sqsupseteq a$. So, by Weak epistemic sufficiency, $\text{str}(a) \leq \text{str}(b)$. \square

Example 9 (Violation of properties for dsp , dps and mean). An argument a with $\text{prec}(a) = 0$ is such that $\text{dsp}(a) = 0$ and $\text{dps}(a) = 1$. Clearly, *neutrality* is violated for dsp and dps . Such an argument also violates *moderation* for dps .

To illustrate other violations we give an example similar to Example 6. Let $\mathbb{K} = \langle \langle \mathcal{V}_p : \{p_1, p_2\}, \mathcal{V}_l : \{q_1, q_2, q_3, q_4\} \rangle, \mathcal{A} : \text{sent}(\mathcal{V}_p), \mathcal{C} : \{(p_1 \wedge p_2 \leftrightarrow q_1), \neg(p_1 \vee p_2) \rightarrow q_2, p_1 \rightarrow q_4\}, \mathbb{P} : \{P\} \rangle$ with the probabilities as in Table 10. We note that $\text{mean}(a_1) > \text{mean}(a_2)$ [resp. $\text{dsp}(a_1) > \text{dsp}(a_2)$] while $\text{dps}(a_2) > \text{dps}(a_1)$ illustrating a violation of *lower compensation* for dps . Since $\text{prec}(a_2) = 0.1 < 1 = \text{prec}(a_1)$ this also gives a counter-example for *precision compensation* and *necessity*, for dps . For a counter-example for *upper compensation* and dsp consider arguments a_3 and a_5 : $\text{dsp}(a_3) < \text{dsp}(a_5)$ and $\text{mean}(a_5) \leq \text{mean}(a_3)$, while $\text{dps}(a_5) < \text{dps}(a_3)$. A counter-example for *strict epistemic sufficiency* and dps is given in view of $\text{dps}(a_2) \neq \text{dps}(a_3)$, although $a_3 \sqsubset a_2$.

Strict epistemic sufficiency and *precision necessity* for dsp is violated in view of hybrid and hybrid' in Example 5, where $\text{dsp}(\text{hybrid}) = \text{dsp}(\text{hybrid}')$ while $\text{hybrid} \sqsubset \text{hybrid}'$ and $\text{prec}(\text{hybrid}) < \text{prec}(\text{hybrid}')$.

Consider $\mathbb{K} = \langle \mathcal{A} : \{p\}, \mathcal{C} : \emptyset, \mathbb{P} : \{P\} \rangle$ where $P(p) = 0.5$ and the arguments $a : \langle \{p\}, p \rangle$ and $b : \langle \emptyset, \top \rangle$. Then $\text{dsp}(a) = 0.5 = \text{dps}(a) = \text{mean}(a)$ and $\text{prec}(a) = 1$, while $\text{dsp}(b) = 0$, $\text{dps}(b) = 1$, $\text{mean}(b) = 0.5$ and $\text{prec}(b) = 0$. The example represents a counter-example for (i) *precision necessity* for $\text{str} \in \{\text{dps}, \text{mean}\}$, (ii) *strict precision sufficiency* for $\text{str} \in \{\text{dps}, \text{mean}\}$ and (iii) *precision sufficiency* for dps .

Example 10 (Violation of properties for precision mean.). In Table 8, we have $\text{dsp}(\text{hybrid}) = 0.7$, $\text{dps}(\text{hybrid}) = 1$, while $\text{precMean}(\text{hybrid}) = 0.595$ (see Table 8). This shows that precMean does not satisfy *domain restriction*. Note that $\text{wave} \sqsubseteq \text{hybrid}$ and $\text{precMean}(\text{wave}) = 0.7$. So, we also have a counter-example for weak and strict *epistemic sufficiency*, as well as for *R-weakening*.

We consider the knowledge base $\mathbb{K} = \langle \langle \mathcal{V}_p : \{p_1, p_2\}, \mathcal{V}_l : \{q_1, q_2, q_3, q_4\} \rangle, \mathcal{A} : \text{sent}(\mathcal{V}_p), \mathcal{C} : \{\neg(p_1 \vee p_2) \rightarrow q_1, (\neg p_2 \vee p_1) \rightarrow q_2, \neg p_2 \rightarrow q_3, (p_1 \wedge p_2) \rightarrow \neg(q_3 \vee q_4), \neg p_1 \wedge p_2 \rightarrow q_4, \neg(p_1 \vee p_2) \rightarrow \neg q_4\}, \mathbb{P} : \{P\} \rangle$ with the probabilities and arguments in Table 11. For a counter-example for *upper* (resp. *lower*) *compensation* consider a_1 and a_2 (resp. a_3). The arguments a_3 and a_5 also provide a counter-example for *precision necessity* since $\text{precMean}(a_3) > \text{precMean}(a_5)$ while $\text{prec}(a_5) = 0.75 > \text{prec}(a_3) = 0.6$.

Example 11 (Violation of lower compensation, Boosted support, and Convex combination). In the knowledge base of Table 11 we have a counter-example for *lower compensation* and bst_m for $m = 1.5$. Note that $\text{bst}_m(a_2) > \text{bst}_m(a_4)$ and $\text{mean}(a_2) \leq \text{mean}(a_4)$ while $\text{dsp}(a_4) > \text{dsp}(a_2)$. In view of Fact 3 the example applies equally to convex_α for $\alpha = 1/3$.

Example 12 (Epistemic Risk Tolerance). We note that, in the example of Table 11, $\text{dps}(a_2) > \text{dps}(a_1)$ [resp. $\text{mean}(a_2) > \text{mean}(a_1)$] (resp. $\text{bst}_{1.5}(a_2) > \text{bst}_{1.5}(a_1)$), while $\text{dsp}(a_2) < \text{dsp}(a_1)$, demonstrating *epistemic risk tolerance* for dps [resp. for mean] (resp. for $\text{bst}_{1.5}$ and $\text{convex}_{2/3}$). For precMean we consider arguments a_1 and a_3 .

3.2. Naively applying argumentation semantics

Argumentation semantics aim at providing a rationale for selecting arguments for acceptance in discursive situations in which arguments and counter-arguments are exchanged. Some requirements are, for instance, that a selection does not contain conflicting arguments, or that a selection is such that any counter-argument to one of its arguments is attacked by some argument in the selection. In this section, we will gradually introduce new notions and observations based on a list of problems. Ultimately the critical discussion will lead to an improved account to be introduced in Section 3.3. In order to define argumentation semantics we first need a notion of argumentative defeat.

Definition 5 (defeat types). Let \mathbb{K} be a knowledge base, str a strength measure, and $a, b \in \text{Arg}(\mathbb{K})$.

rebuttal: a rebuts b if (1) $\text{str}(a) \geq \text{str}(b)$ and (2) $\text{Con}(a) \vdash_C \neg \text{Con}(b)$.

TABLE 10 Arguments and state space for the knowledge base $\mathbb{K} = \langle \langle \mathcal{V}_p : \{p_1, p_2\}, \mathcal{V}_l : \{q_1, q_2, q_3, q_4\} \rangle, \mathcal{A} : \text{sent}(\mathcal{V}_p), \mathcal{C} : \{(p_1 \wedge p_2 \leftrightarrow q_1), \neg(p_1 \vee p_2) \rightarrow q_2, p_1 \rightarrow q_4\}, \mathbb{P} : \{P\} \rangle$, Example 9.

State	p_1	p_2	P	q_1	q_2	q_3	q_4	Argument	dsp	dps	mean	precMean
s_1	0	0	0.1	0	1	\diamond	\diamond	$a_1 : \langle \{p_1 \wedge p_2\}, q_1 \rangle$	0.7	0.7	0.7	0.7
s_2	0	1	0.1	0	\diamond	\diamond	\diamond	$a_2 : \langle \{\neg(p_1 \vee p_2)\}, q_2 \rangle$	0.1	1	0.55	0.055
s_3	1	0	0.1	0	\diamond	\diamond	1	$a_3 : \langle \emptyset, q_3 \rangle$	0	1	0.5	0
s_4	1	1	0.7	1	\diamond	\diamond	1	$a_4 : \langle \{p_1\}, q_4 \rangle$	0.8	1	0.9	0.72
								$a_5 : \langle \{\neg(p_1 \wedge p_2)\}, \neg q_1 \rangle$	0.3	0.3	0.3	0.3

TABLE 11 Arguments and the state space for $\mathbb{K} = \langle \langle \mathcal{V}_p : \{p_1, p_2\}, \mathcal{V}_l : \{q_1, q_2, q_3, q_4\} \rangle, \mathcal{A} : \text{sent}(\mathcal{V}_p), \mathcal{C} : \{\neg(p_1 \vee p_2) \rightarrow q_1, (\neg p_2 \vee p_1) \rightarrow q_2, \neg p_2 \rightarrow q_3, (p_1 \wedge p_2) \rightarrow \neg(q_3 \vee q_4), \neg p_1 \wedge p_2 \rightarrow q_4, \neg(p_1 \vee p_2) \rightarrow \neg q_4\}, \mathbb{P} : \{P\} \rangle$ (see Example 10).

	p_1	p_2	P	q_1	q_2	q_3	q_4	Argument	dsp	dps	mean	precMean	bst _{1.5}	bst _{2.5}
s_1	0	0	0.1	1	1	1	0	$a_1 : \langle \{p_1\}, p_1 \rangle$	0.5	0.5	0.5	0.5	0.5	0.5
s_2	0	1	0.4	\diamond	\diamond	\diamond	1	$a_2 : \langle \{\neg(p_1 \vee p_2)\}, q_1 \rangle$	0.1	1	0.55	0.055	0.7	0.46
s_3	1	0	0.25	\diamond	1	1	\diamond	$a_3 : \langle \{\neg p_2 \vee p_1\}, q_2 \rangle$	0.6	1	0.8	0.48	0.867	0.76
s_4	1	1	0.25	\diamond	1	0	0	$a_4 : \langle \{\neg p_2\}, q_3 \rangle$	0.35	0.75	0.55	0.33	0.62	0.51
								$a_5 : \langle \{\neg p_1 \wedge p_2\}, q_4 \rangle$	0.4	0.65	0.525	0.394	0.567	0.5

undercut: a undercuts b if (1) $\text{str}(a) \geq \text{str}(b)$ and (2) $\text{Con}(a) \vdash_C \neg \bigwedge \text{Sup}'$ for $\emptyset \neq \text{Sup}' \subseteq \text{Sup}(b)$.

undercut': a undercuts' b if (1) $\text{str}(a) \geq \text{str}(@(\text{Sup}(b)))$, where $@(\text{Sup}(b)) = \langle \text{Sup}(b), \bigwedge \text{Sup}(b) \rangle$, and (2) $\text{Con}(a) \vdash_C \neg \bigwedge \text{Sup}'$ for $\emptyset \neq \text{Sup}' \subseteq \text{Sup}(b)$.

Lemma 1. Suppose Weak Epistemic Sufficiency holds for str . Let $a, b \in \text{Arg}(\mathbb{K})$.

1. If $\text{Con}(a) \vdash_C \text{Con}(b)$ and $\text{Sup}(a) = \text{Sup}(b)$ then $\text{str}(a) \leq \text{str}(b)$.
2. $\text{str}(@(\text{Sup}(a))) \leq \text{str}(a)$.
3. If $\text{Sup}(a) \subseteq \text{Sup}(b)$ then $\text{str}(@(\text{Sup}(a))) \geq \text{str}(@(\text{Sup}(b)))$.
4. If a undercuts b , a also undercuts' b .

Proof: *Ad 1.* Suppose $\text{Con}(a) \vdash_C \text{Con}(b)$ and $\text{Sup}(a) = \text{Sup}(b)$. By Fact 2, $\text{dsp}(a) = \text{dsp}(b)$ and $\text{dps}(b) \geq \text{dps}(a)$. By Weak Epistemic Sufficiency, $\text{str}(a) \leq \text{str}(b)$. *Ad 2.* This is a special case of item 1 since $\text{Sup}(a) = \text{Sup}(@(\text{Sup}(a)))$ and $\text{Con}(@(\text{Sup}(a))) \vdash_C \text{Con}(a)$. *Ad 3.* In this case $\text{Con}(@(\text{Sup}(b))) \vdash_C \text{Con}(@(\text{Sup}(a)))$. By Fact 2, $\text{dps}(a) \geq \text{dps}(b)$ and $\text{dsp}(a) \geq \text{dsp}(b)$. By Weak Epistemic Sufficiency, $\text{str}(@(\text{Sup}(a))) \geq \text{str}(@(\text{Sup}(b)))$. *Ad 4.* Suppose a undercuts b . In order to show that a undercuts' b we only have to show that $\text{str}(a) \geq \text{str}(@(\text{Sup}(b)))$. Since $\text{str}(a) \geq \text{str}(b)$ this follows with Item 2. \square

We are now in a position to define argumentation frameworks and subsequently argumentation semantics.

Definition 6 (AF). An *argumentation framework* based on a knowledge base \mathbb{K} is a pair $\langle \text{Arg}(\mathbb{K}), \text{def} \rangle$ where def is a (non-empty) set of defeat-types (as in Definition 5) for a given measure of argument strength str .

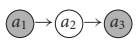
3.2.1. Problem 1. Reinstatement and threshold selections

When selecting arguments for acceptance one may follow a naive “threshold-strategy” according to which one sets a threshold τ , say $\tau = 0.55$, and simply selects all arguments which are stronger than τ (e.g., according to their degree of support, or another argument strength measure).⁸ This strategy, however, leads to various kinds of problems. One, illustrated in Example 16 below, is that following this strategy one may be left of with arguments whose conclusions form an inconsistent set. In this sense, the strategy selects too many arguments. On the other hand, this strategy does not validate a central principle from argumentation theory: reinstatement. It states that an argument which is defended by a set of accepted arguments, should also be accepted.

Example 13 (Reinstatement). Consider the following knowledge base: $\mathbb{K} = \langle \langle \mathcal{V}_p : \{w_1, w_2, w_3\}, \mathcal{V}_l : \{\text{scene}, \text{home}, \text{pub}, \dots\} \rangle, \mathcal{A} : \{w_1, w_2, w_3\}, \mathcal{C} : \{w_1 \rightarrow \text{home}, w_2 \rightarrow \text{scene}, w_3 \rightarrow \text{pub}, \neg(\text{scene} \wedge \text{home}), \neg(\text{scene} \wedge \text{pub}), \text{pub} \rightarrow \text{home}\}, \mathbb{P} : \{P\} \rangle$. In our scenario we have 3 witnesses. Witness 1 states, among other things, that Mr. X was in his home town at the time of the murder ($w_1 \rightarrow \text{home}$), witness 2 states that Mr. X was at the scene of the murder ($w_2 \rightarrow \text{scene}$), and witness 3 that he was at the pub ($w_3 \rightarrow \text{pub}$). Mr. X cannot be both at the scene and in his home town. Also, the pub is located in Mr. X's home town. Witness 1 has a reliability of 0.82 (e.g., we estimate that she tells the truth in 82/100 cases), witness 2 of 0.91 and witness 3 of 0.6. After correcting the probabilities according to the states consistent with \mathcal{C} (see Section 2.4) we obtain the ones listed in Table 12. There, we also state three key arguments a_1, a_2 and a_3 , their strength and the argumentation framework based on $\text{str} = \text{mean}$ and rebuttal.

⁸ For instance, Hunter (2013) uses a threshold of 0.5 to define his *epistemic extensions*.

TABLE 12 State space, probabilities, and arguments for Example 13.

State	w_1	w_2	w_3	$P(s_i)$	$P^C(s_i)$	home	scene	pub	Argument	[dsp, dps]	mean
s_1	0	0	0	$0.18 \cdot 0.09 \cdot 0.4$	0.044	\diamond	\diamond	\diamond	a_1	[0.506, 0.551]	0.528
s_2	0	1	0	$0.18 \cdot 0.91 \cdot 0.4$	0.449	0	1	0	a_2	[0.449, 0.494]	0.472
s_3	1	0	0	$0.82 \cdot 0.09 \cdot 0.4$	0.202	1	0	\diamond	a_3	[0.304, 0.551]	0.427
s_4	1	0	1	$0.82 \cdot 0.09 \cdot 0.6$	0.304	1	0	1			
AF											

We omit states that are incompatible with C . We calculate $P^C(s_i)$ by $P(s_i)/e$ where $e = P(\|T\|_C) = \sum_{i=1}^4 P(s_i) = 0.146$ (see Section 2.4). We have the arguments $a_1 : \{\{w_1\}, \text{pub}\}$, $a_2 : \{\{w_2\}, \text{scene}\}$ and $a_3 : \{\{w_3\}, \text{home}\}$. The argumentation framework on the right (bottom) is based on rebuttal and $\text{str} = \text{mean}$.

The strongest argument a_1 is in favor of Mr. X being in his home town, which would clear Mr. X from suspicion. If we subscribe to this argument, however, the argument a_3 for him being in the pub becomes quite reasonable, since its only attacker a_2 (him being at the scene) is refuted. If we put ourselves in the investigative spirit of a detective working the case, it seems reasonable to select arguments a_1 and a_3 to form an investigative and/or explanatory hypothesis (despite the strength of a_3 being below a threshold such as 0.5, both in terms of its degree of support or mean value). This hypothesis, may then lead us to the decision to investigate the question whether he was at the pub further in order to either substantiate or refute our stance.

Observation 2 (Reinstatement). Argumentative reinstatement is not validated in naive threshold-based approaches for selecting arguments. This motivates other types of selections, since in specific reasoning scenarios, such as the formation of explanatory hypothesis, reinstatement is a reasonable argumentative principle.

Since threshold-based selection does not allow for reinstatement we will also study other selection types, in particular those introduced by Dung (1995) for abstract argumentation.

Definition 7 (Argumentation Semantics, Dung, 1995). Given an $\text{AF} = \langle \text{Arg}(\mathbb{K}), \text{Def} \rangle$ and a set of arguments $\mathcal{E} \subseteq \text{Arg}(\mathbb{K})$ we define

- \mathcal{E} is *conflict-free* iff $(\mathcal{E} \times \mathcal{E}) \cap \text{Def} = \emptyset$.
- \mathcal{E} *defends* some $a \in \text{Arg}(\mathbb{K})$ iff for every defeater b of a there is a $c \in \mathcal{E}$ that defeats b .
- \mathcal{E} is *admissible* iff \mathcal{E} is conflict-free and it defends every $a \in \mathcal{E}$.
- \mathcal{E} is *complete* iff \mathcal{E} is admissible and it contains every $a \in \text{Arg}(\mathbb{K})$ it defends.
- \mathcal{E} is *grounded* iff it is the unique \subseteq -minimal complete extension.
- \mathcal{E} is α -*accepted* in case $\mathcal{E} = \{a \in \text{Arg}(\mathbb{K}) \mid \text{str}(a) > \alpha\}$ (where $\alpha \in [0, 1]$, typically $\alpha = 0.5$).
- \mathcal{E} is *preferred* iff \mathcal{E} is a \subseteq -maximal complete extension.
- \mathcal{E} is *stable* iff \mathcal{E} is conflict-free and $\mathcal{E} \cap \text{Arg}(\mathbb{K})$ defeats every $a \in \text{Arg}(\mathbb{K}) \setminus \mathcal{E}$.

In the remainder of this section we show that naively applying these semantics to AFs leads to various problems. In the next

section we present an alternative approach to resolve (some of) these issues.

Let us first highlight differences between the two types of defeat, rebut and undercut.

3.2.2. Selecting arguments with inconsistent support with some semantics

Example 14 (The possibility of inconsistent supports.). This example is similar to Example 3, where $\mathbb{K} = \langle \mathcal{A} : \{p, \neg p\}, \mathcal{C} : \{p \rightarrow q\}, \mathbb{P} : \{P\} \rangle$, except for the probability function P which is adjusted as described in Table 13 (left). On the right hand of the figure we describe the arguments and their respective strengths. It seems clear that the argument $a_{\neg p}$ in favor of $\neg p$ is preferable to the argument a_p in favor of p , given that $P(\| \neg p \|) = 4/7 > 3/7 = P(\| p \|)$. What about a_q in favor of q , though? On the one hand, it is based on the assumption p , since only if p we know for certain that q . On the other hand, it comes with HOU, since for the case $\neg p$ we are under-informed about q : q is possible (and so is $\neg q$). Thus, q seems to have more in its favor than $\neg q$ and a reasoner committing to q seems not irrational, possibly even so when also selecting $a_{\neg p}$ and therefore committing to $\neg p$. Note that such a reasoner will not be committed to an inconsistent set of arguments (since $\{\neg p, q\}$ is consistent). In the following, we will show how some attack types support this kind of selections, while others do not.

Observation 3 (Possibility of inconsistent supports.). In probabilistic argumentation, when situations of HOU arise, we can rationalize selections of arguments with mutually inconsistent supports (but consistent conclusions).

It should be mentioned, though, that this observation is normatively cautious. We do not claim that from a normative philosophical perspective such selections are permissible (although they may be), we merely claim that agents are in a position to rationalize such selections. A formal framework modeling such selections is therefore at least attractive from a descriptive cognitive perspective (but possibly also from a normative philosophical perspective).

Let us now consider the different defeat-types in combination with our various argument strength measures, in order to see how they model the example. The underlying argumentation frameworks are illustrated in Table 13 (right, bottom).

Rebut. The argumentation framework based on rebut is in conformity with the rationale underlying Observation 3. Despite

TABLE 13 The state space, probabilities (left), arguments, and argumentation frameworks for different attack forms (right), Example 14 for $\mathbb{K} = \langle \mathcal{A} : \{p, \neg p\}, \mathcal{C} : \{p \rightarrow q\}, \mathbb{P} : \{P\} \rangle$.

				Argument	Mean(a)	$\text{bst}_{m < 4}(a)$ (= $\text{convex}_{\alpha < 2/3}$)	$\text{bst}_{m > 4}(a)$ (= $\text{convex}_{\alpha > 2/3}$)
				$a_p : \langle \{p\}, p \rangle$	$3/7$	$3/7$	$3/7$
				$a_{\overline{p}} : \langle \{\neg p\}, \neg p \rangle$	$4/7$	$4/7$	$4/7$
State	p	q	P	$a_q = \langle \{p\}, q \rangle$	$5/7$	$> 4/7$	$< 4/7$
s_1	1	1	$3/7$	$a_{p \vee \overline{q}} : \langle \{p\}, p \vee \neg q \rangle$	$5/7$	$> 4/7$	$< 4/7$
s_2	0	\Diamond	$4/7$	$a_{p \vee q} : \langle \{p\}, p \vee q \rangle$	$5/7$	$> 4/7$	$< 4/7$
				rebut			
				undercut			
				undercut'			

the fact that a_q is based on the support p and $a_{\bar{p}}$ attacks a_p , $a_{\bar{p}}$, and a_q are selected.

Undercut. In case $\text{str}(a_q) > \text{str}(a_{\bar{p}})$ (e.g., where $\text{str} = \text{mean}$ or $\text{str} = \text{bst}_m$ with $m < 4$, see Table 13), the latter is not sufficiently strong to defeat a_q leading to a selection analogous to the one based on rebut. Conceptually, however, undercut creates a tension in this and similar examples. While the rationale underlying undercut is that arguments with inconsistent supports should not both end up in the same selection, in our example they do since $a_{\bar{p}}$ is not strong enough to undercut a_q (while condition (2) of Definition 5 is met, condition (1) is not, which renders undercut unsuccessful in this case). This incoherence is resolved with our variant undercut'.

Undercut'. In contrast to undercut, for undercut' for $a_{\bar{p}}$ to attack a_q it merely needs to be at least as strong as a_p . Therefore, in all scenarios the attack is successful (see right column in Table 13). Therefore, undercut' does not allow for a selection of arguments with mutually inconsistent supports (We prove this impossibility in Section 3.4 after solving some other problems.).

3.2.3. Problem 2: selecting arguments with inconsistent conclusions with rebut

When *only* working with rebut, we run into problems with inconsistent arguments, as the following example shows.

Example 15 (Inconsistent conclusions with rebut.). Consider $\mathbb{K} = \langle \langle \mathcal{V}_p : \{p\}, \mathcal{V}_l : \{q\} \rangle, \mathcal{A} : \{p, \neg p\}, \mathcal{C} : \emptyset, \mathbb{P} : \{P\} \rangle$ where $P(p) = 0.5$. We have, for instance, the following arguments: $a_{\top} = \langle \emptyset, \neg(p \wedge \neg p) \rangle$, $a_p = \langle \{p\}, p \rangle$, $a_{\bar{p}} = \langle \{\neg p\}, \neg p \rangle$, $a_q = \langle \{p, \neg p\}, q \rangle$ and $a_{\bar{q}} = \langle \{p, \neg p\}, \neg q \rangle$. In an approach based on rebut, we get, for instance, a complete extension \mathcal{E} containing the arguments a_{\top} , a_p and a_q . The latter argument, or any argument for q based on \mathbb{K} , is problematic in that it is based on an inconsistent support. Rebut does not effectively filter out such arguments. We also note that $[\text{dsp}(a_{\top}), \text{dps}(a_{\top})] = \{1\}$ while $[\text{dsp}(a_q), \text{dps}(a_q)] = [0, 1] = [\text{dsp}(a_{\bar{q}}), \text{dps}(a_{\bar{q}})]$. So, for any strength measure respecting Domain Restriction, $\text{str}(a_{\top}) \geq \text{str}(a_q) = \text{str}(a_{\bar{q}})$ and so a_{\top} undercuts a_q and $a_{\bar{q}}$. This shows that with undercut-based attacks inconsistent arguments are “automatically” filtered out.

In order to deal with the problem of inconsistent arguments when using rebuts, we can either manually sort out inconsistent arguments (as proposed in Wu and Podlaszewski, 2014), or use inconsistency-undercuts (as proposed in Arieli and Straßer, 2020) in addition to rebuts.

Inconsistency Undercut. Where $a, b \in \text{Arg}(\mathbb{K})$, $\text{Sup}(b) \vdash_{\mathcal{C}} \perp$, $a = \langle \emptyset, \neg \bigwedge \text{Sup}(b) \rangle$ inconsistency-undercuts b .

Lemma 2. Let str satisfy Domain restriction. If a inconsistency undercuts b , then (i) a undercuts [resp. undercuts'] b , (ii) $\text{str}(a) = 1$, and (iii) there is no argument that defeats a (according to rebut, undercut, undercut', or inconsistency undercut).




Proof: Suppose a inconsistency undercuts b . Since $\text{Sup}(a) = \emptyset$, by Domain restriction, $\text{str}(a) = \text{dsp}(a) = \text{dps}(a) = \inf_{P \in \mathbb{P}} P(\| \top \|_{\mathcal{C}}) = \sup_{P \in \mathbb{P}} P(\| \top \|_{\mathcal{C}}) = 1$. This is (ii). For (i) it is sufficient to show that $\text{str}(a) \geq \text{str}(b)$. This follows trivially from (ii). For (iii) assume toward a contradiction that some c defeats a . Since $\text{Sup}(a) = \emptyset$, this cannot be an undercut, undercut', or inconsistency undercut. Suppose c rebuts a . So, $\text{Con}(c) \vdash_{\mathcal{C}} \bigwedge \text{Sup}(b)$ and therefore $\text{Con}(a) \vdash_{\mathcal{C}} \neg \text{Con}(c)$. Moreover, $\emptyset \vdash_{\mathcal{C}} \neg \text{Sup}(c)$. So, $\text{dsp}(c) = \text{dps}(c) = 0$ since $\| \text{Sup}(c) \|_{\mathcal{C}} = \| \diamond \text{Con}(c) \|_{\mathcal{C}} = \emptyset$. Therefore, $\text{str}(c) < \text{str}(a)$, a contradiction. \square

3.2.4. Problem 3: ($n > 2$)-conflicts and selecting arguments with inconsistent conclusions

The following example illustrates that even in scenarios with exclusively precise probabilities (so, all arguments have precision 1) all discussed types of attack lead to problems.

Example 16 (($n > 2$)-conflicts and inconsistent selections.). Let $\mathbb{K} = \langle \langle \mathcal{V}_p : \{p_1, p_2\}, \mathcal{V}_l : \emptyset \rangle, \mathcal{A} : \emptyset(\Gamma) \setminus \Gamma, \mathcal{C} : \emptyset, \mathbb{P} : \{P\} \rangle$ where $\Gamma = \{p_1, p_2, \neg(p_1 \wedge p_2)\}$, P is given in Table 14 (right). There we also list arguments (left) with their corresponding strengths and an excerpt of the underlying argumentation framework (center), relative to any of the defeat-types, rebut, undercut and undercut'. As the reader can easily verify, there is a complete extension (highlighted) containing a_1, a_2 , and a_n . The problem with this selection is that it contains inconsistent conclusions, namely p_1, p_2 , and $\neg(p_1 \wedge p_2)$. The same problem occurs with α -selections for, e.g., $\alpha \leq 0.54$.

TABLE 14 The state space, probabilities (right), arguments and argumentation framework (left) for $\mathbb{K} = (\langle \mathcal{V}_p : \{p_1, p_2\}, \mathcal{V}_l : \emptyset \rangle, \mathcal{A} : \wp(\Gamma) \setminus \Gamma, \mathcal{C} : \emptyset, \mathbb{P} : \{P\})$, $\Gamma = \{p_1, p_2, \neg(p_1 \wedge p_2)\}$, and any of the defined strength measures str, Example 16.

Argument	str	Attack diagram	State	p_1	p_2	P
$a_1 = \langle \{p_1\}, p_1 \rangle$	0.55	0.55				
$a_2 = \langle \{p_2\}, p_2 \rangle$	0.55	0.55	 s_1	0	0	0.05
$a_n = \langle \{\neg(p_1 \wedge p_2)\}, \neg(p_1 \wedge p_2) \rangle$	0.85	0.85	 s_2	0	1	0.40
$a_b = \langle \{p_1, p_2\}, p_1 \wedge p_2 \rangle$	0.15	0.15	 s_3	1	0	0.40
$a_{\neg} = \langle \{p_2, \neg(p_1 \wedge p_2)\}, \neg p_1 \rangle$	0.40	0.425		1	1	0.15
$a_{\neg} = \langle \{p_1, \neg(p_1 \wedge p_2)\}, \neg p_2 \rangle$	0.40	0.425				

In simple scenarios such as the one above, one may reasonably expect a reasoner to make a consistent selection of arguments.⁹

Observation 4 (Inconsistency with regular AFs). Naively applying argumentation semantics in the context of probabilistic argumentation may lead to inconsistent selections, even for simple scenarios only including two probabilistic variables and no higher-order uncertainties. We consider this a serious problem, which we try to accommodate in the next section.

3.3. Using hyper-arguments: a refined method for argument selection

Given a knowledge based \mathbb{K} , in order to enforce the consistency of the set of conclusion of a given complete extension we will make use of what we call *hyper-arguments* (collected in the set $\text{HArg}(\mathbb{K})$, see Definition 8 below), i.e., arguments written as $[a_1, \dots, a_n]$ (where $a_1, \dots, a_n \in \text{Arg}(\mathbb{K})$). Hyper-arguments express the idea that if one were to accept each a_1, \dots, a_n then one cannot accept a regular argument $b \in \text{Arg}(\mathbb{K})$ for which $\{a_1, \dots, a_n, b\}$ is conflicting. For this a specific type of hyper-argument based defeat, so-called h-defeats, are introduced. From the argumentation theoretic perspective hyper-defeats express the meta-argumentative consideration that a reasoner should not commit to an inconsistent set of arguments. Therefore, hyper-arguments do not contribute to the content-level of a discussion, but rather they express constraints on argument selection.

In the following, we will make this idea formally precise, illustrate it with examples and study meta-theoretic properties in

Section 3.4. As we will see, working both with normal and hyper-arguments, as well as both with h-defeats and defeats, suffices to ensure the consistency of the set of conclusions of complete extensions (and some other properties) and therefore avoids the problem pointed out in Observation 4.

Definition 8 (Hyper-arguments.). Given a knowledge base \mathbb{K} and $a_1, \dots, a_n \in \text{Arg}(\mathbb{K})$, $[a_1, \dots, a_n]$ is a hyper-argument (based on \mathbb{K}). We call a_1, \dots, a_n the *components* of $[a_1, \dots, a_n]$. We let $\text{Sup}([a_1, \dots, a_n]) = \bigcup_{i=1}^n \text{Sup}(a_i)$ and $\text{Con}([a_1, \dots, a_n]) = \bigwedge_{i=1}^n \text{Con}(a_i)$. We denote by $\text{HArg}(\mathbb{K})$ the set of all hyper-arguments a based on \mathbb{K} .

In the following we will use the convention to use sub-scripted variables a_i, b_i , etc. for regular arguments (in $\text{Arg}(\mathbb{K})$) and non-subscripted variables a, b , etc. for both regular arguments and hyper-arguments. We use ‘argument’ as a generic term covering both regular and hyper-arguments.

Attacks are generalized to the level of hyper-arguments by letting, for instance, $[a_1, \dots, a_n]$ *h-rebut* b in case $\text{Con}([a_1, \dots, a_n]) \vdash_{\mathcal{C}} \neg \text{Con}(b)$. A hyper-argument is defeated resp. h-defeated by another regular argument resp. hyper-argument if one of its component arguments a_i is defeated resp. h-defeated (see Definition 9 below). While defeat is a relation on the domain $\text{Arg}(\mathbb{K}) \times (\text{Arg}(\mathbb{K}) \cup \text{HArg}(\mathbb{K}))$, h-defeat is a relation on the domain $\text{HArg}(\mathbb{K}) \times (\text{HArg}(\mathbb{K}) \cup \text{Arg}(\mathbb{K}))$.

Definition 9 (h-defeat.). Let \mathbb{K} be a knowledge base. *h-defeats* define a relation on $\text{HArg}(\mathbb{K}) \times (\text{Arg}(\mathbb{K}) \cup \text{HArg}(\mathbb{K}))$. Let $a = [a_1, \dots, a_n], b = [b_1, \dots, b_m] \in \text{HArg}(\mathbb{K})$ and $c \in \text{Arg}(\mathbb{K})$.

- a *h-rebuts* c iff $\text{Con}(a) \vdash_{\mathcal{C}} \neg \text{Con}(c)$.
- a *h-rebuts* b iff there is an $i \in \{1, \dots, m\}$ for which a h-rebuts b_i .
- a *h-undercuts* c iff $\text{Con}(a) \vdash_{\mathcal{C}} \neg \bigwedge \text{Sup}(c)$.
- a *h-undercuts* b iff for some $i \in \{1, \dots, m\}$, a h-undercuts b_i .

Note that unlike regular defeats, h-defeats do not consider argument strength. The reason is that h-defeats encode meta-argumentative considerations concerning the consistency of selections of arguments. For such considerations, argument strength is of no concern.

Definition 10 (Regular defeats). Let \mathbb{K} be a knowledge base. *Defeats* define a relation on $\text{Arg}(\mathbb{K}) \times (\text{Arg}(\mathbb{K}) \cup \text{HArg}(\mathbb{K}))$ where the part on $\text{Arg}(\mathbb{K}) \times \text{Arg}(\mathbb{K})$ is defined as in Definition 5, and

⁹ However, there may be limitations to the requirement of consistency. As is well-known from cases such as the lottery paradox (Kyburg, 1961) or the preface paradox (Makinson, 1965), complex scenarios may give rise to inconsistent belief states, possibly even for rational reasoners. Although this is, as the reader may expect, a deep philosophical problem, see Douven and Williamson (2006) for a critical discussion. Clearly, though, in the simple examples included in our paper it should be considered irrational to hold inconsistent beliefs and we also don’t expect it to be descriptively adequate. A discussion of the mentioned paradoxical scenarios in the context of the formalism presented in this paper is left for future occasions.

TABLE 15 A list of argumentation properties.

Property	Definition
Component closure	$[a_1, \dots, a_n] \in \mathcal{E}$ iff $a_1, \dots, a_n \in \mathcal{E}$.
Direct consistency	If $a_1, a_2 \in \mathcal{E}$ then $\text{Con}(a_1), \text{Con}(a_2) \not\vdash_C \perp$.
Indirect consistency	If $a_1, \dots, a_n \in \mathcal{E}$ then $\text{Con}(a_1), \dots, \text{Con}(a_n) \not\vdash_C \perp$.
Weakening	If $a_1 \in \mathcal{E}$ and $\text{Con}(a_1) \vdash_C \phi$ then also $(\text{Sup}(a_1), \phi) \in \mathcal{E}$.
Support consistency	If $a_1, \dots, a_n \in \mathcal{E}$ then $\bigcup_{i=1}^n \text{Sup}(a_i) \not\vdash_C \perp$.
Dir. support closure	If $a_1 \in \mathcal{E}$ then for every a_2 for which $\text{Sup}(a_2) \subseteq \text{Sup}(a_1)$, $a_2 \in \mathcal{E}$.
Ind. support closure	If $a_1, \dots, a_n \in \mathcal{E}$ and $b_1 \in \text{Arg}(\mathbb{A}\mathbb{F})$ is s. t. $\text{Sup}(b_1) \subseteq \bigcup_{i=1}^n \text{Sup}(a_i)$, then $b_1 \in \mathcal{E}$.
Logical closure	If $a_1, \dots, a_n \in \mathcal{E}$, $(\bigcup_{i=1}^n \text{Sup}(a_i), \phi) \in \mathcal{E}$ for all ϕ for which $\bigcup_{i=1}^n \text{Sup}(a_i) \vdash_C \phi$.

A property \mathbb{P} in the left column of the table holds for an argumentation semantics sem in case for every (hyper) argumentation framework $\mathbb{A}\mathbb{F}$ and every sem -extension of $\mathbb{A}\mathbb{F}$ the right column holds.

some $a \in \text{Arg}(\mathbb{K})$ rebuts [resp. undercuts, undercuts'] some $b = [b_1, \dots, b_n] \in \text{HArg}(\mathbb{K})$ iff a rebuts [resp. undercuts, undercuts'] some component b_i of b .

Fact 5. Let \mathbb{K} be a knowledge base, $a \in \text{Arg}(\mathbb{K}) \cup \text{HArg}(\mathbb{K})$ and $b \in \text{HArg}(\mathbb{K})$. a defeats [resp. h-defeats] b (according to rebut, undercut, undercut' and consistency undercut) iff a defeats [resp. h-defeats] some component b_i of b .

Having defined regular and hyper-arguments and different notions of defeat among them, we are now in a position to generalize our notion of argumentation frameworks to include hyper-arguments.

Definition 11 (Hyper AF, h-AF). A *hyper-argumentation framework* based on a knowledge base \mathbb{K} is a pair $(\langle \text{Arg}(\mathbb{K}), \text{HArg}(\mathbb{K}) \rangle, \langle \text{Def}, \text{Hdef} \rangle)$ where Def is a relation of regular defeat and Hdef a relation of hyper-defeat based on rebut and/or undercut and/or undercut' and/or inconsistency undercut.

In the remainder, we consider three types of frameworks:

- (1) *rebut-based h-AFs*, where $\text{Def} = \{\text{rebut}, \text{cons.undercut}\}$ and $\text{Hdef} = \{\text{h-rebut}\}$
- (2) *undercut-based h-AFs*, where $\text{Def} = \{\text{undercut}\}$ and $\text{Hdef} = \{\text{h-undercut}\}$
- (3) *undercut'-based h-AFs*, where $\text{Def} = \{\text{undercut}'\}$ and $\text{Hdef} = \{\text{h-undercut}\}$

Argumentation semantics are adjusted to the case with hyper-arguments as expected. We only need to adjust the notion of defense: defeats need to be counter-defeated, while h-defeats need to be counter-h-defeated.

Definition 12 (Argumentation Semantics). Given an h-AF $\mathbb{A}\mathbb{F} = (\langle \text{Arg}(\mathbb{K}), \text{HArg}(\mathbb{K}) \rangle, \langle \text{Def}, \text{Hdef} \rangle)$ and a set of arguments $\mathcal{E} \subseteq \text{Arg}(\mathbb{K}) \cup \text{HArg}(\mathbb{K})$ we say

- \mathcal{E} is *conflict-free* iff $(\mathcal{E} \times \mathcal{E}) \cap (\text{Def} \cup \text{Hdef}) = \emptyset$.

- \mathcal{E} *defends* some $a \in \text{Arg}(\mathbb{K}) \cup \text{HArg}(\mathbb{K})$ iff for every defeater [resp. h-defeater] b of a there is a $c \in \mathcal{E}$ that defeats [resp. h-defeats] b .
- \mathcal{E} is *admissible* iff \mathcal{E} is conflict-free and it defends every $a \in \mathcal{E}$.
- \mathcal{E} is *complete* iff \mathcal{E} is admissible and it contains every $a \in \text{Arg}(\mathbb{K}) \cup \text{HArg}(\mathbb{K})$ it defends.
- \mathcal{E} is *preferred* iff \mathcal{E} is a \subseteq -maximal complete extension.
- \mathcal{E} is *stable* iff \mathcal{E} is conflict-free and $\mathcal{E} \cap \text{Arg}(\mathbb{K})$ defeats every $a \in \text{Arg}(\mathbb{K}) \setminus \mathcal{E}$.

Our definition requires that only h-defeats can defend from h-defeats. In [Appendix C.1 \(Supplementary material\)](#), we show that allowing regular defeats to defend from h-defeats leads to the same complete extensions (see Proposition 18).

Example 17. Let $\mathbb{K} = \langle \mathcal{A} : \{p, \neg p\}, \mathcal{C} : \emptyset, \mathbb{P} : \{P\} \rangle$ where $P(\|p\|) = 0.6$ (and $P(\|\neg p\|) = 0.4$). Let $a_p = \langle \{p\}, p \rangle$, $a_{\neg p} = \langle \{\neg p\}, \neg p \rangle$. Let defeat be rebut (or undercut). In [Figure 2](#) (left), we see an excerpt of an hyper-argumentation framework based on \mathbb{K} . With the above definitions there is a slight redundancy in that every regular argument a has a hyper-argument $[a]$ as counter-part. Note that a_p is defended from the hyper-attack by $[a_{\neg p}]$ by its hyper-argument counterpart $[a_p]$. Unlike $a_{\neg p}$ and $[a_{\neg p}]$, a_p and $[a_p]$ are part of the unique preferred extension. Note that $a_{\neg p}$ and $[a_{\neg p}]$ cannot be defended from the defeat by a_p .

In the following examples we will omit hyper-argumentative counterparts of regular arguments in the attack diagrams. For instance, [Figure 2](#) (left) will be simplified to [Figure 2](#) (center). In [Appendix C.2](#) (Proposition 19), we show that it is possible to work without hyper-arguments of the form $[a]$, i.e., to identify them with their regular counterparts. In our example this variant also results in [Figure 2](#) (right).

We have omitted the grounded extension from Definition 12. Example 17 illustrates why. While we would expect a_p to be contained in the grounded extension, it is not since it is in need to be defended from the h-defeat by $[a_{\neg p}]$, but no non-attacked argument is able to do so. So, in many cases the grounded extension will not be informative since it will only contain arguments without h-attackers (e.g., those with tautological conclusions).¹⁰

Example 18 (Example 16 cont.). In [Figure 3](#), we show excerpts of the argumentation frameworks for Example 16, now enriched with hyper-arguments. We have three preferred extensions, $\mathcal{E}_1 = \{a_1, a_n, [a_1, a_n], \dots\}$ (left), $\mathcal{E}_2 = \{a_2, a_n, [a_2, a_n], \dots\}$ (center) and $\mathcal{E}_3 = \{a_1, a_2, [a_1, a_2], \dots\}$ (right). We note that the problematic complete extension from Example 16, including arguments a_1, a_2 and a_n is not anymore admissible in the setup with hyper-arguments. One of the reasons is that the h-defeat from $[a_2, a_n]$ on a_1 cannot be defended. Indeed, the defeat from $[a_2, a_n]$ expresses the consistency constraint that if we accept a_2 and a_n then we shall

¹⁰ A similar observation can be made for stable semantics. While these need not exist in frameworks with odd defeat-cycles in regular AFs, the situation worsens in hyper-argumentation frameworks due to the presence of h-defeats (see Example 18). We include stable semantics nevertheless in Definition 12 since they satisfy some rationality postulates that don't hold for preferred semantics (see [Table 16](#)).

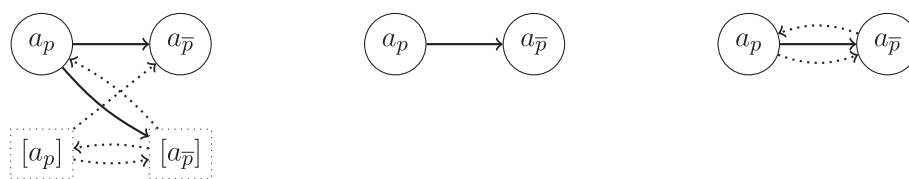


FIGURE 2

Illustration for Example 17. Dotted arrows indicate h-defeats, solid arrows regular defeats. **(Left)** Detailed presentation. **(Center)** Compact presentation omitting simple hyper-arguments. **(Right)** The presentation obtained by the variant defined in Appendix C.2 (Supplementary material).

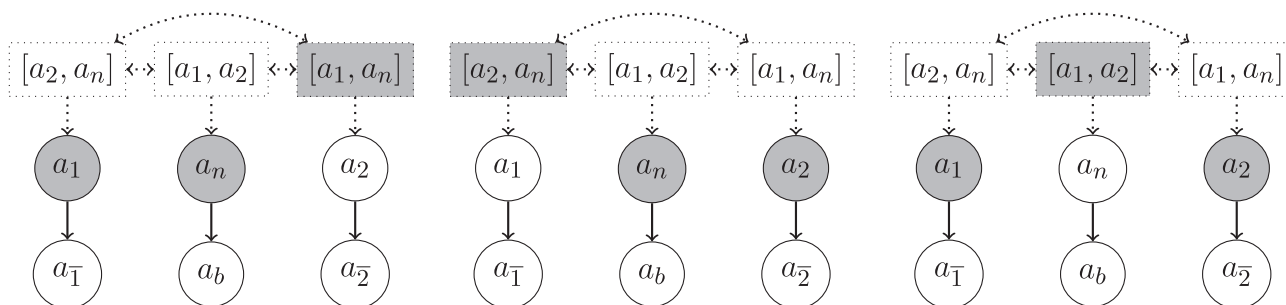


FIGURE 3

Excerpt of the hyper-argumentation frameworks for Example 18 with rebut, undercut, resp. undercut'. Highlighted are three preferred extensions (from left to right). As mentioned above, for the sake of compactness of presentation we identified simple hyper-arguments (i.e., hyper-arguments with only one component) with their component and omitted symmetric h-defeats between arguments whenever there are regular defeats present between their components (e.g., the h-defeat between $[a_1]$ and $[a_2]$ is omitted since a_1 is defeated by a_2).

not accept a_1 . We also note that neither of these three extensions is stable, e.g., $a_2 \notin \mathcal{E}_1$ and a_2 is also not defeated by \mathcal{E}_1 .

3.4. Rationality postulates for hyper-argumentation frameworks

We now study meta-theoretic properties of hyper-argumentation frameworks. Table 15 contains various properties, often called “rationality postulates” (see Caminada and Amgoud, 2007; Arieli et al., 2021). We will investigate these for our different attack types and for argument strength measures that satisfy Weak Epistemic Sufficiency and Domain Restriction. We consider two general scenarios: a *naive* one without hyperarguments (as discussed in Section 3.2) and a *hyper* one with hyperarguments. Table 16 provides an overview of our results. Proofs are provided in Appendix B (Supplementary material). We summarize:

Observation 5 (Key observations.). Our results show that hyper-argument based probabilistic argumentation satisfies the desiderata discussed in Observations 1, 3, and 4.

Concerning Observation 1 we employ argument strength measures that satisfy weak epistemic sufficiency to do justice to the intuition that an argument such as *hybrid* is stronger than an argument such as *wave* due to the presence of higher-order uncertainty.

In order to model the intuition underlying Observation 3 one may use hyper-argumentation frameworks based on rebuttals: in such frameworks both $a_{\bar{p}}$ and a_q can be present in

the same complete extension, without causing an inconsistent conclusion set.

Finally, we overcome the problem of the existence of complete extensions with inconsistent conclusion sets identified for regular argumentation frameworks in Observation 4: all of the studied hyper-argumentation frameworks satisfy the postulate of Indirect Consistency.

In the remainder of this section we illustrate the lack of some properties from Table 16 with examples. For this we first take another look at Example 14, this time with hyper-arguments.

Example 19 (Example 14 cont.). In Figure 4, we show excerpts of the argumentation frameworks for Example 14 for rebut (left), undercut (center), and undercut' (right), now enriched with hyper-arguments. In each figure we highlight a preferred extension. We note that the one on the right is unique.

Example 20 (Counter-examples, Rationality Postulates.). In Figures 3, 4, we observe the following violations of rationality postulates.

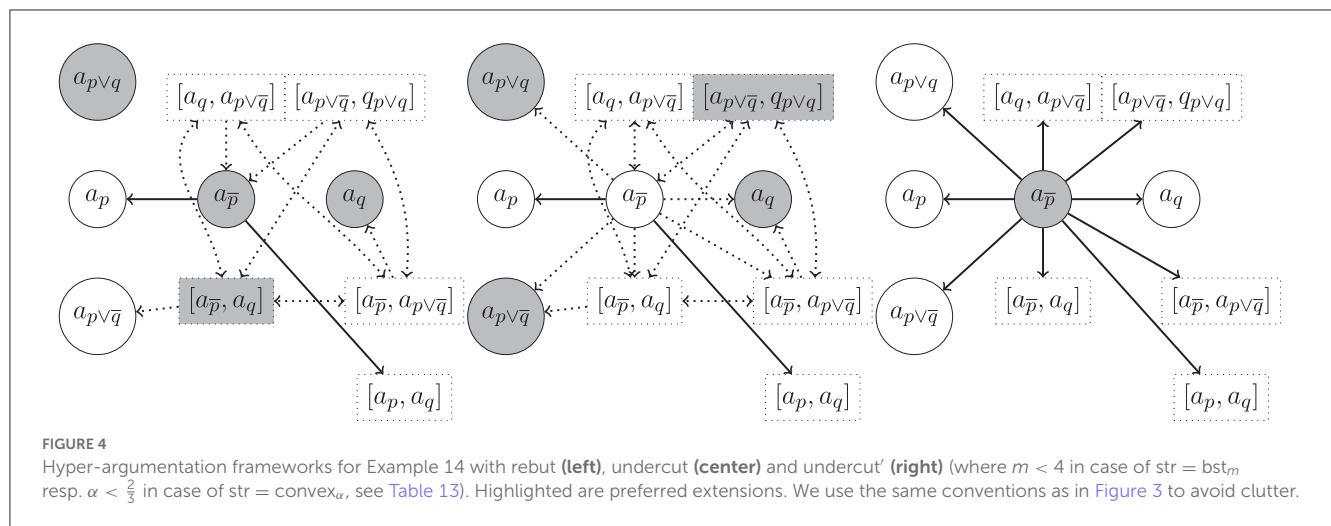
Support Consistency. Support consistency is violated for rebut in Figure 4 (left). There both $a_{\bar{p}}$ and a_q are contained in the given preferred extension, although $\text{Sup}(a_{\bar{p}}), \text{Sup}(a_q) \vdash_C \perp$.

Logical Closure. Logical closure is violated for rebut, undercut and undercut' for complete and preferred extensions, as can be seen in Figure 3 (left). Although a_1 and a_n are accepted in the given preferred extension, the argument $a_2 = \langle \{p_1, \neg(p_1 \wedge p_2)\}, \neg p_2 \rangle$ is not (it is and cannot be defended from the defeat by a_2 since by Definition 12 a defense from a regular defeat must be in terms of a

TABLE 16 Overview: rationality postulates.

Method Arguments Attack form(s)	Naive Arg(\mathbb{K}) Rebut cons. u.cut	Naive Arg(\mathbb{K}) Undercut (cons. u.cut)	Naive Arg(\mathbb{K}) Undercut' (cons. u.cut)	Hyper HArg(\mathbb{K}) Rebut cons. u.cut	Hyper HArg(\mathbb{K}) Undercut (cons. u.cut)	Hyper HArg(\mathbb{K}) Undercut' (cons. u.cut)
Component closure	n.a.	n.a.	n.a.	✓ [Corollary 1]	✓ [Corollary 1]	✓ [Corollary 1]
Direct consistency	✓ [Proposition 11]	✓ [Proposition 11]	✓ [Proposition 11]	✓ [Proposition 11]	✓ [Proposition 11]	✓ [Proposition 11]
Indirect consistency	✗ [Example 16]	✗ [Example 16]	✗ [Example 16]	✓ [Proposition 12]	✓ [Proposition 13]	✓ [Proposition 13]
Weakening	✓ [Proposition 10]	✓ [Proposition 9]	✓ [Proposition 9]	✓ [Proposition 10]	✓ [Proposition 9]	✓ [Proposition 9]
Support consistency	✗ [Example 16]	✗ [Example 16]	✗ [Example 16]	✗ [†]	✓ [Proposition 15]	✓ [Proposition 15]
Logical closure	✗ [Example 16]	✗ [Example 16]	✗ [Example 16]	✗ [†]	✗ [†]	✓* [Proposition 17]
Dir. support closure	✗ [Example 14]	✗ [Example 14]	✗ [Example 14]	✗ [†]	✗ [†]	✓ [Proposition 14]
Ind. support closure	✗ [Example 16]	✗ [Example 16]	✗ [†]	✗ [†]	✗ [†]	✓* [Proposition 16]

All propositions and corollaries are to be found in [Appendix B \(Supplementary material\)](#). Properties marked by ✓* apply only to stable semantics. Counter-examples for other semantics can be found in Example 20, as well as counter-examples for properties marked with [†].



regular defeat and therefore the hyper-defeat of $[a_1, a_n]$ on a_2 is not sufficient to defend a_2 from a_2 .

Support Closure. Direct support closure is violated for both rebuts and undercuts. For rebuts we have in the preferred extension of [Figure 4](#) (left), a_q selected, but not a_p although $\text{Sup}(a_p) = \text{Sup}(a_q)$. Similarly for undercuts, in [Figure 4](#) (center). The violation of indirect support closure is an immediate consequence.

As for undercut' and *indirect support closure* we consider [Figure 3](#) (right): although a_1 and a_2 are selected, $a_b = \langle \{p_1, p_2\}, p_1 \wedge p_2 \rangle$ is not since it cannot be defended from the undercut' from a_n . Note that the h-undercut on a_n by the selected $[a_1, a_2]$ is not sufficient to defend a_b from a regular undercut': Definition 12 requires a defense from a regular defeat in terms of a regular defeat (i.e., an undercut' in this case).

4. Empirical study

In this section, we discuss a small empirical study we conducted on evaluating argument strength in the context of higher-order

uncertainty.¹¹ Our main objective was to investigate the following research questions:

RQ1. Is argument evaluation more context-sensitive than our logical model predicts? To answer this question we consider two reasoning contexts: an abstract one where participants have to reason about the probability to draw balls from an urn, and one practical medical context. In both scenarios, the participants face arguments of the same underlying logical form in our representation (see [Appendices D, E](#) in [Supplementary material](#) for details) but with different informal interpretations. For such arguments our model calculates the same degrees of support and possibilities, and therefore it predicts the same argument strengths. Similarly, we want to know whether across different contexts arguments of the same logical form are evaluated equally by our participants.

RQ2. How do the different argument strength measures from Section 3.1 predict the participants' answers? In particular, which values of the parameters m for bst_m resp. α for convex_α are

¹¹ We leave empirical studies concerning argument selection/semantics (Section 3) for a future occasion.

empirically adequate (possibly relative to fixed reasoning contexts, see RQ1)?

RQ3. Which rationality postulates from Section 3.4 are met resp. violated by the participants' answers? In particular, is the intuition behind our Observation 1 empirically adequate?

The study was conducted in the context of three university seminars on the Bachelor and Master level of philosophy programs. Altogether 42 students participated. The questionnaire encompasses 19 questions and is structured into 3 reasoning scenarios. Each scenario comes with a number of arguments built on the basis of the available information. For each argument, the participants were asked to rate its strength in a scale with 10 subdivisions, reaching from *very weak* to *very strong* (see Figure 5 for two arguments in the context of the second scenario). We list all scenarios of the questionnaire in detail in Appendix C (Supplementary material).¹²

The three reasoning scenarios covered by the questionnaire are: (S1) one of the well-known Ellsberg scenarios (Ellsberg, 1961, see Example 2), (S2.1) a less abstract re-phrasing of the Ellsberg scenario in terms of a medical investigation, (S2.2) a variant of (S2.1) in which more emphasis is given to imprecise probabilistic information, as discussed in Section 2.3 (similar to Example 6).

Table 17 gives an overview on our results. We now evaluate our findings.

4.1. Concerning RQ1

We first observe that the reasoning context is crucial for the assessment of argument strength. We note that scenarios 1 and 2.1 have the same formal structure and therefore our model predicts the same argument strength assessments for arguments of the same logical form (indicated by $\alpha, \beta, \gamma, \delta$ and ϵ in Table 17). Indeed, within scenario 2.1, the evaluation of the strength of arguments of the same logical form (Q10 and Q12 resp. Q11 and Q13) remained relatively stable (max. variance is 0.02 between the mean values) among our participants. However, if we compare arguments of the same logical form between scenario 1 and scenario 2 we see clear differences. For instance for arguments of type α we have a difference of 0.13 in the mean, for arguments of type ϵ a difference of .02. In particular, the evaluation of α in the context of Q1 is 0.45 and in the context of Q9 it is 0.32. What is also striking is that for imprecise arguments there is basically no variance between the two scenarios. This asymmetry is surprising and we don't have an explanation for it.

4.2. Concerning RQ2

When averaging over all questions the optimal value for m is ≈ 2.05 and the one for α is ≈ 0.51 . In view of this the mean measure is a good approximation of the empirical results. However, when zooming into the different types of arguments we observe that the

m (resp. α) value is contextual, depending on where the $[dsp, dps]$ -interval is situated. With Table 18 we observe the tendency that m grows the more the weight of the $[dsp, dps]$ -interval moves toward 1. This means that the reasoning becomes more cautious resp. risk averse in such cases. For instance, the average strength estimation of arguments of type γ with $[dsp, dps] = [1/3, 1]$ is 0.58 (closer to the dsp), while the average strength estimation of arguments of type μ with $[dsp, dps] = [0, 1/3]$ is 0.29 (closer to the dps).

4.3. Concerning RQ3

Epistemic sufficiency could be generally verified in the study.¹³ This reflects positively on our Observation 1 which can be considered empirically verified in view of our small study. Participants show typically *Risk tolerant* reasoning and therefore violated *Risk aversion*. *Upper compensation* could not convincingly be verified in our questionnaire. *Strict precision sufficiency*, *Precision necessity* and *compensation* only fare slightly better. In contrast, the acceptance rates for *Lower compensation* and for *Precision sufficiency* are in average high.

Before moving to *Domain restriction* and *Moderation*, we make two methodological remarks of caution. First, the scale of the questionnaire was not numerical and therefore it does not directly represent the interval $[0, 1]$ in which our technical notions such as dsp, dps , etc. are measured. Therefore, a validation of criteria such as Domain restriction based on this questionnaire has to be interpreted with caution, since we naively mapped the interval in questionnaire to the interval $[0, 1]$ (preserving scaling). Second, we interpreted the answers of the participants charitably, e.g., when evaluating Domain restriction we checked if the answer is "roughly" within the corresponding interval. Despite these methodological hurdles we consider the empirical study informative also for these criteria since it allows us to see discrepancies between the replies concerning logically equivalent arguments (indicated by types α, \dots, ϵ in Table 17) in different settings. We observe that *Domain Restriction* is violated for the types α and δ (even under a charitable interpretation of the answers). Interestingly for the imprecise arguments (so, arguments for which $dsp(a) < dps(a)$) *Domain Restriction* could be empirically verified. It is again the precise arguments as opposed to the imprecise ones, for which we find violations of *Moderation*. We see some divergence for arguments of type α between to two scenarios (in S2.1 and S2.2 *Moderation* is verified for α -type arguments, not in S1), while for arguments of type δ *Moderation* fails in more than 50% in both scenarios. One explanation may be

¹³ We note that the only case with a low acceptance rate is Scenario 2.1 for Weak epistemic sufficiency (34%). However, this value is due to the fact that there are 12 different pairwise argument combinations (e.g., Q8 and Q9, Q10 and Q11, etc.) in which participants could violate the criterion. The 34% feature only participants who validated weak epistemic sufficiency for every pair. If we consider each such pair separately, the acceptance rate per pair is rather high. In most of the pairs we see a very high acceptance rate for Weak epistemic sufficiency, only when comparing Q11 and Q13 resp. Q10 and Q12 our participants struggle. The reason is that for these particular pairs Weak epistemic sufficiency would demand equal strength attribution.

¹² In this study, we did not randomize the order of presentation for each participant in order to avoid priming.

TABLE 17 Overview on the results from the empirical study on argument strength.

Scenario	scenario S1 (Ellsberg)					Scenario 2.1 (Medical)							S2.2 (Impr. Prob.)		
Question	Q1	Q2	Q3	Q4	Q5	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17
Argument type	α	β	γ	δ	ϵ	δ	α	γ	β	γ	β	ϵ	α	λ	μ
dsp	1/3	0	1/3	2/3	1	2/3	1/3	1/3	0	1/3	0	1	1/3	1/3	0
dps	1/3	2/3	1	2/3	1	2/3	1/3	1	2/3	1	2/3	1	1/3	2/3	1/3
Claim type	at	at	\vee	\vee	\vee	\vee	at	\vee	at	\vee	at	\vee	at	at	at
Average Strength	0.45	0.4	0.6	0.69	0.97	0.7	0.32	0.58	0.34	0.56	0.35	0.99	0.33	0.56	0.29
ordering	$\beta < \alpha < \gamma < \delta < \epsilon$					$\alpha(<)\beta < \gamma < \delta < \epsilon$							$\mu < \alpha < \lambda$		
Domain restr.	0.31	0.98	1	0.38	0.67	0.37	0.51	0.98	0.95	0.95	0.88	0.80	0.53	0.67	0.7
Moderation	0.36	0.36	0.64	0.45	-	0.41	0.76	0.78	0.68	0.78	0.61	-	0.76	0.5	0.29
Weak Ep. suff.	0.6					0.34							0.73		
Str. Ep. suff.	0.57					0.73							0.68		
Ep. risk tolerance	0.74					0.66							0.32		
Upper compensation	0.52		0.5			0.24	0.54	0.24	0.54	0.24	0.54		0.97		1
Lower compensation	0.79		0.88			0.88	0.63	0.88	0.63	0.88	0.63		0.97		1
Precision suff.	0.79		0.88			0.88	0.63	0.88	0.63	0.88	0.63		-		
Str. precision suff.	0.48		0.5			0.63	0.38	0.63	0.38	0.63	0.38		-		
Precision nec.	0.48		0.5			0.63	0.38	0.63	0.38	0.63	0.38		0.03		
Prec. compensation	0.48		0.5			0.63	0.38	0.63	0.38	0.63	0.38		0.03		1

Listed are the different scenarios (S1, S2.1, and S2.2) and their respective questions. Each question is concerned with the evaluation of the strength of an argument presented in an informal way. Some of these arguments share the same type w.r.t. their degrees of support and possibility (indicated by α, \dots, ϵ). Precise arguments (i.e., arguments a for which $\text{dsp}(a) = \text{dps}(a)$) are underlined. The exact logical form is presented in [Appendix E \(Supplementary material\)](#). We also list the type of claim (atomic “at” vs. disjunctive “ \vee ”). Below we list the average strength assessment of the participants and an empirical evaluation of the properties from Section 3.1. In the first block we present properties concerned with arguments in isolation (Domain Restriction and Moderation, but note our cautious remarks concerning the evaluation of the survey with respect to these criteria in the main text). The second block concerns properties where arguments are compared. In these cases we analyse the two main scenarios S1 and S2 separately. For the desiderata weak/strong epistemic sufficiency and epistemic risk tolerance we compared the answers block-wise according to the scenarios 1, 2.1, and 2.2., i.e., 60% of participants validated weak epistemic sufficiency for all questions in scenario 1. For the rest of the criteria we picked out paradigmatic pairings of arguments. A number below two questions indicates that the arguments corresponding to those two questions were compared to each other according to the desideratum, e.g., the first value 0.52 for upper compensation means that w.r.t. the arguments in Q1 and Q2 52% of the participants answered in accordance with upper compensation. In scenario 2.1 whenever questions have the same number, it means those were compared to each other according to the desideratum, i.e., 24% of participants fulfilled upper compensation when comparing questions 8, 10, and 12 to each other. The hyphen ‘-’ indicates that the criterion is not applicable.

TABLE 18 Optimal m values for different argument types.

Type	[dsp, dps]	Optimal m	Average strength
μ	[0, 1/3]	1.15	0.29
λ	[1/3, 2/3]	1.46	0.56
β	[0, 2/3]	1.84	0.36
γ	[1/3, 1]	2.63	0.58

Third, we show how abstract argumentation semantics ([Dung, 1995](#)) can be applied to our framework given different (standard) notions of attack (versions of undercut and rebut). It is well-known from deductive argumentation that violations of rationality postulates can occur if one proceeds too naively. We proposed a solution based on *hyper-arguments*, which express consistency constraints. Given that our framework generalizes deductive argumentation and Hunter’s probabilistic argumentation, the solution applies also there. In the context of probabilistic

argumentation Dung’s semantics are rarely applied. For example, [Haenni \(2009\)](#) does not propose any rationale for selecting arguments for selection, while [Hunter \(2013\)](#) uses threshold semantics. We consider Dung’s semantics attractive for several reasons. First, they are widely applied and well-researched in formal argumentation ([Baroni et al., 2018](#)); second, being based on notions such as conflict-freeness and defendability, they are very intuitive; and third, they allow for reinstatement, a principle that is not (in general) validated by threshold semantics. The latter is in particular interesting when generating explanatory hypotheses (see Example 13 and Observation 2). In this context we note that it is sometimes distinguished between an epistemic and a constellations approach ([Hunter, 2012](#)). While in the former probabilities express a doxastic attitude toward arguments, in the latter they express how likely it is that arguments belong to and/or are relevant to a certain discursive situation. Our approach clearly belongs in the epistemic camp. We note that the interpretation of argument strength and defeat in structured non-probabilistic argumentation seems more in line

with the epistemic approach and it is where reinstatement is often applied.¹⁴

The paper presents only a first step to systematically integrate reasoning with HOU in abstract argumentation. In future work, we intend to enhance the empirical study, both in terms of the number of participants and also in scope, by a stronger focus on the impact of context on argument strength, and by including questions of argument selection (e.g., is reinstatement used by participants when generating hypotheses and explanations?, etc.). Another application of our framework is to study in more detail reasoning in the context of multiple agents (e.g., considering testimony, higher-order evidence, and dialogue¹⁵). According to (Elkin and Wheeler, 2016; Elkin, 2021; Henderson, 2021) situations of peer disagreements and/or where higher-order evidence matters (e.g., evidence provided by expert panels, etc.) should not be modeled by naively aggregating beliefs, since this may overstate precision, but it should be modeled in terms of credal sets, i.e., in terms of HOU. Our framework provides some of the basic ingredients to model argumentation in such contexts. In the present work we restricted the focus on purely epistemic reasoning by not considering other practical utilities. A possible enhancement of our study is to widen the focus and incorporate decision theories under HOU (such as Gilboa and Schmeidler, 2004).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

¹⁴ Also in the context of the constellations approach imprecise probabilities have been utilized (see Fazzinga et al., 2022). It is particularly useful in cases in which our knowledge concerning the probabilities or arguments and attacks is limited.

¹⁵ See Caminada (2018) for an overview on dialogical methods and abstract argumentation semantics.

References

- Arieli, O., Borg, A., and Straßer, C. (2021). "Characterizations and classifications of argumentative entailments," in *Proceedings of KR* (Montreal), 52–62.
- Arieli, O., and Straßer, C. (2015). Sequent-based logical argumentation. *Argument Comput.* 6, 73–99. doi: 10.1080/19462166.2014.1002536
- Arieli, O., and Straßer, C. (2020). "On minimality and consistency tolerance in logical argumentation frameworks," in *Computational Models of Argument* (IOS Press), 91–102.
- Baroni, P., Caminada, M., and Giacomin, M. (2018). Abstract argumentation frameworks and their semantics. *Handb. Formal Argument.* 1, 157–234.
- Beirlaen, M., Heyninck, J., and Straßer, C. (2018). "A critical assessment of Pollock's work on logic-based argumentation with suppositions," in *Proceedings of the NMR* (Tempe, AZ), 63–72.
- Besnard, P., and Hunter, A. (2001). A logic-based theory of deductive arguments. *Artif. Intell.* 128, 203–235. doi: 10.1016/S0004-3702(01)00071-6
- Besnard, P., and Hunter, A. (2018). A review of argumentation based on deductive arguments. *Handb. Formal Argument.* 1, 435–482.
- Bradley, S. (2019). "Imprecise probabilities," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Metaphysics Research Lab, Stanford University).

Author contributions

All concepts and insights in the paper are a result of discussions and joint research between CS and LM. While it is hard to entangle individual contributions, CS's emphasis was on the theoretical foundation including meta-proofs, while LM's emphasis was on gathering and evaluating empirical data. All authors contributed to the article and approved the submitted version.

Funding

LM has been supported by a student stipend from the Gerhard C. Starck Foundation.

Acknowledgments

We would like to thank the reviewers. One of the reports was particularly in-depth, thorough and very helpful to improve the quality of this paper: thanks for that! Finally, thanks to all the students who participated in our survey.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2023.1133998/full#supplementary-material>

- Brevers, D., Cleeremans, A., Goudriaan, A. E., Bechara, A., Kornreich, C., Verbanck, P., et al. (2012). Decision making under ambiguity but not under risk is related to problem gambling severity. *Psychiatry Res.* 200, 568–574. doi: 10.1016/j.psychres.2012.03.053
- Caminada, M. (2018). “Argumentation Semantics as Formal Discussion,” in *Handbook of Formal Argumentation*, Vol. 1, eds P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre (College Publications), 487–518.
- Caminada, M., and Amgoud, L. (2007). On the evaluation of argumentation formalisms. *Artif. Intell.* 171, 286–310. doi: 10.1016/j.artint.2007.02.003
- Cramer, M., and Dauphin, J. (2019). A structured argumentation framework for modeling debates in the formal sciences. *J. Gen. Philos. Sci.* 51, 219–241. doi: 10.1007/s10838-019-09443-z
- Cramer, M., and Dietz Saldanha, E.-A. (2020). “Logic programming, argumentation and human reasoning,” in *Logic and Argumentation*, eds M. Dastani, H. Dong, and L. van der Torre (Cham: Springer International Publishing), 58–79.
- De Groot, K., and Thurik, R. (2018). Disentangling risk and uncertainty: when risk-taking measures are not about risk. *Front. Psychol.* 9, 2194. doi: 10.3389/fpsyg.2018.02194
- Douven, I. (2010). Simulating peer disagreements. *Stud. History Philos. Sci. A* 41, 148–157. doi: 10.1016/j.shpsa.2010.03.010
- Douven, I., and Williamson, T. (2006). Generalizing the lottery paradox. *Brit. J. Philos. Sci.* 57, 755–779. doi: 10.1093/bjps/axl022
- Dung, P., Kowalski, R., and Toni, F. (2009). Assumption-based argumentation. *Argument. Artif. Intell.* 199–218. doi: 10.1007/978-0-387-98197-0_10
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77, 321–358. doi: 10.1016/0004-3702(94)00041-X
- Elkin, L. (2021). The precautionary principle and expert disagreement. *Erkenntnis*. doi: 10.1007/s10670-021-00457-y
- Elkin, L., and Wheeler, G. (2016). Resolving peer disagreements through imprecise probabilities. *Noûs* 52, 260–278. doi: 10.1111/nous.12143
- Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Q. J. Econ.* 75, 643–669.
- Euteneuer, F., Schaefer, F., Stuermer, R., Boucsein, W., Timmermann, L., Barbe, M. T., et al. (2009). Dissociation of decision-making under ambiguity and decision-making under risk in patients with Parkinson’s disease: a neuropsychological and psychophysiological study. *Neuropsychologia* 47, 2882–2890. doi: 10.1016/j.neuropsychologia.2009.06.014
- Fazzinga, B., Flesca, S., and Furfaro, F. (2022). “Abstract argumentation frameworks with marginal probabilities,” in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence* (Vienna).
- Gilboa, I., and Schmeidler, D. (2004). “Maxmin expected utility with non-unique prior,” in *J. Math. Econ.* 18, 141–153.
- Haenni, R. (2009). Probabilistic argumentation. *J. Appl. Logic* 7, 155–176. doi: 10.1016/j.jal.2007.11.006
- Hahn, U. (2020). Argument quality in real world argumentation. *Trends Cogn. Sci.* 24, 363–374. doi: 10.1016/j.tics.2020.01.004
- Hahn, U., and Oaksford, M. (2007). The rationality of informal argumentation: a Bayesian approach to reasoning fallacies. *Psychol. Rev.* 114, 704–732. doi: 10.1037/0033-295X.114.3.704
- Henderson, L. (2021). Higher-order evidence and losing one’s conviction. *Noûs*. 56, 513–529. doi: 10.1111/nous.12367
- Hunter, A. (2012). Some foundations for probabilistic abstract argumentation. *Comma* 245, 117–128. doi: 10.3233/978-1-61499-111-3-117
- Hunter, A. (2013). A probabilistic approach to modelling uncertain logical arguments. *Int. J. Approximate Reason.* 54, 47–81. doi: 10.1016/j.ijar.2012.08.003
- Hunter, A. (2022). Argument strength in probabilistic argumentation based on defeasible rules. *Int. J. Approximate Reason.* 146, 79–105. doi: 10.1016/j.ijar.2022.04.003
- Hunter, A., Polberg, S., and Thimm, M. (2020). Epistemic graphs for representing and reasoning with positive and negative influences of arguments. *Artif. Intell.* 281, 103236. doi: 10.1016/j.artint.2020.103236
- Hunter, A., and Thimm, M. (2017). Probabilistic reasoning with abstract argumentation frameworks. *J. Artif. Intell. Res.* 59, 565–611. doi: 10.1613/jair.5393
- Josang, A. (2001). A logic for uncertain probabilities. *Int. J. Uncertainty Fuzziness Knowledge Based Syst.* 9, 279–311. doi: 10.1142/S0218488501000831
- Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263.
- Kyburg, H. E. (1961). *Probability and the Logic of Rational Belief*. Middletown, CT: Wesleyan University Press.
- Makinson, D. C. (1965). The paradox of the preface. *Analysis* 25, 205–207.
- Mercier, H., and Sperber, D. (2017). *The Enigma of Reason*. Cambridge: Harvard University Press.
- Modgil, S., and Prakken, H. (2014). Theaspic+framework for structured argumentation: a tutorial. *Argument Comput.* 5, 31–62. doi: 10.1080/19462166.2013.869766
- O’Donnell, R. (2021). Keynes and knight: risk-uncertainty distinctions, priority, coherence and change. *Cambridge J. Econ.* 45, 1127–1144. doi: 10.1093/cje/beab034
- Oren, N., Norman, T. J., and Preece, A. (2007). Subjective logic and arguing with evidence. *Artif. Intell.* 171, 838–854. doi: 10.1016/j.artint.2007.04.006
- Pfeifer, N. (2013). *On Argument Strength*. Dordrecht: Springer.
- Pfeifer, N., and Pankka, H. (2017). Modeling the ellsberg paradox by argument strength. *arXiv preprint arXiv:1703.03233*. doi: 10.48550/arXiv.1703.03233
- Saldanha, E.-A. D., and Kakas, A. (2019). Cognitive argumentation for human syllogistic reasoning. *Künstliche Intelligenz* 33, 229–242. doi: 10.1007/s13218-019-00608-y
- Santini, F., Josang, A., and Pini, M. S. (2018). “Are my arguments trustworthy? abstract argumentation with subjective logic,” in *2018 21st International Conference on Information Fusion (FUSION)* (Cambridge).
- Savage, L. J. (1972). *The Foundations of Statistics*. New York, NY: Dover Publications.
- Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.
- Trotzke, P., Starcke, K., Pedersen, A., Müller, A., and Brand, M. (2015). Impaired decision making under ambiguity but not under risk in individuals with pathological buying-behavioral and psychophysiological evidence. *Psychiatry Res.* 229, 551–558. doi: 10.1016/j.psychres.2015.05.043
- Wu, Y., and Podlaszewski, M. (2014). Implementing crash-resistance and non-interference in logic-based argumentation. *J. Logic Comput.* 25, 303–333. doi: 10.1093/logcom/exu017
- Zhang, L., Dong, Y., Ji, Y., Tao, R., Chen, X., Ye, J., et al. (2015). Trait-related decision making impairment in obsessive-compulsive disorder: evidence from decision making under ambiguity but not decision making under risk. *Sci. Rep.* 5, 17312. doi: 10.1038/srep17312



OPEN ACCESS

EDITED BY

Loizos Michael,
Open University of Cyprus, Cyprus

REVIEWED BY

Pietro Baroni,
University of Brescia, Italy
Khalid Al-Khatib,
University of Groningen, Netherlands

*CORRESPONDENCE

Antonis Bikakis
✉ a.bikakis@ucl.ac.uk

RECEIVED 14 December 2022

ACCEPTED 02 June 2023

PUBLISHED 16 June 2023

CITATION

Bikakis A, Flouris G, Patkos T and Plexousakis D
(2023) Sketching the vision of the Web of
Debates. *Front. Artif. Intell.* 6:1124045.
doi: 10.3389/frai.2023.1124045

COPYRIGHT

© 2023 Bikakis, Flouris, Patkos and Plexousakis.
This is an open-access article distributed under
the terms of the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Sketching the vision of the Web of Debates

Antonis Bikakis^{1*}, Giorgos Flouris², Theodore Patkos² and
Dimitris Plexousakis²

¹Department of Information Studies, University College London, London, United Kingdom, ²Institute of
Computer Science, Foundation for Research and Technology–Hellas, Heraklion, Greece

The exchange of comments, opinions, and arguments in blogs, forums, social media, wikis, and review websites has transformed the Web into a modern agora, a virtual place where all types of debates take place. This wealth of information remains mostly unexploited: due to its textual form, such information is difficult to automatically process and analyse in order to validate, evaluate, compare, combine with other types of information and make it actionable. Recent research in Machine Learning, Natural Language Processing, and Computational Argumentation has provided some solutions, which still cannot fully capture important aspects of online debates, such as various forms of unsound reasoning, arguments that do not follow a standard structure, information that is not explicitly expressed, and non-logical argumentation methods. Tackling these challenges would give immense added-value, as it would allow searching for, navigating through and analyzing online opinions and arguments, obtaining a better picture of the various debates for a well-intentioned user. Ultimately, it may lead to increased participation of Web users in democratic, dialogical interchange of arguments, more informed decisions by professionals and decision-makers, as well as to an easier identification of biased, misleading, or deceptive arguments. This paper presents the vision of the Web of Debates, a more human-centered version of the Web, which aims to unlock the potential of the abundance of argumentative information that currently exists online, offering its users a new generation of argument-based web services and tools that are tailored to their real needs.

KEYWORDS

online debate analysis, computational argumentation, computational persuasion, web technologies, human-centered AI

1. Introduction

From the plain publishing of content¹ to the collaborative contribution of knowledge through social media² and the annotation of content with machine-processable semantic information,³ the Web has been constantly reshaping. The development of the Social Web (the social aspect of Web 2.0) has brought about a significant change in the way people use the Web. Nowadays, people around the world access the Web to rate a hotel or a restaurant; they share comments on the story and the writing style of a book; they use it to like or dislike a photograph, a video, or the whole lifework of a music band; they write opinions in blogs; they discuss subjects of any matter in forums; they substantiate opinions in wikis citing

1 https://en.wikipedia.org/wiki/Web_2.0#Web_1.0

2 http://en.wikipedia.org/wiki/Web_2.0

3 https://en.wikipedia.org/wiki/Semantic_Web

sources of diverse reliability. Currently, the Web is flooded with opinions and arguments touching topics related to just about everything important or insignificant that happens or has happened or may happen in our world.

Unfortunately, all these colorful, diverse, contradictory, interesting or indifferent opinions get lost; *scripta manent*, yet opinions are currently not uploaded as machine-processable data, they are not interlinked, and it is extremely difficult for Web users to find opinions and arguments related to a particular subject, let alone to evaluate them, characterize them based on objective or subjective criteria, or select the ones that would appeal more to them. Current search engines can only help the user access the pages containing arguments on a topic; manual effort is then required for making sense out of the multitude of contradictory and diverse results returned, for identifying the relations among the available arguments and supportive data, or for analyzing their credibility.

Building on the recent advancements in *Machine Learning*, *Natural Language Processing*, and *Computational Argumentation*, there have been some attempts to unlock the potential of this information. These include an ontology for representing arguments using well-defined, structured formats (Rahwan et al., 2007), methods for argument mining (Stede and Schneider, 2018; Lawrence and Reed, 2019), software tools for argument analysis and visualization (Reed et al., 2017), argument search engines (Wachsmuth et al., 2017b; Stab et al., 2018; Chen et al., 2019), persuasive chatbots (Chalaguine and Hunter, 2020), and autonomous debating systems (Slonim et al., 2021). However, existing efforts fall short in two ways: first, there is still no mature technology allowing the reliable extraction of arguments from text for annotation and further automated processing; second, there are still no general models for realistic arguments, which would be able to capture all aspects of our everyday argumentative dialogues or debates on topics of general concern, such as global warming, international politics, or the energy crisis. Especially since Dung's seminal paper on Abstract Argumentation Frameworks (Dung, 1995), we have developed a very good understanding of the relation between argumentation and logic-based reasoning. However, human dialogues and debates often involve arguments based on *implicit information* (e.g., commonsense knowledge), may resort to *unsound reasoning* (e.g., proof-by-example), or employ *non-logical argumentation methods* (e.g., peer-pressure, use of emotionally loaded arguments, authoritative claims). The study of such aspects and, more generally, the study of the *ethos* (appeal to the credibility of the speaker) and *pathos* (appeal to the emotions of the audience) of argumentation, is not yet as mature as the study of the *logos* of argumentation, in the context of Artificial Intelligence.

Furthermore, online arguments and opinions are not just put forward to be heard, but they have a *purpose* and their processing needs to be purposeful as well. There is, therefore, a need for a new generation of Web tools that will assist humans in reaching conclusions using arguments that are not only formally structured, but are also tailored to the particular characteristics of the *audience* that they are addressed to and the *context* in which they are made, in order to be better comprehensible, more relevant and, therefore, more effective. For any topic, it is important to provide Web users with an overview of all different viewpoints; it is equally important, however, the presentation of these viewpoints to take into account

the background knowledge and cognitive characteristics of each individual user.

To address these challenges and needs, we propose and sketch the design of a new version of the Web, which we call the *Web of Debates*. Its ultimate goal will be to offer the means for assisting humans in participating in debates and collective decision making processes with well-justified and persuasive arguments, as well as in identifying biased, misleading or deceptive arguments. It will be a *global, human-centric AI system*, which, taking advantage of advanced AI methods, will be able to process and analyse the huge amount of natural language arguments and opinions that are available online, and provide its users with personalized, user-friendly services for retrieving, filtering, evaluating and visualizing this information, helping them better make sense of the different viewpoints, draw their own conclusions and take informed decisions about any matter of personal or public concern. The aims of this paper are to describe this vision, identify the requirements and challenges of its realization, discuss the theoretical and technological advancements that are needed to address them, and provide a roadmap toward its realization. Another aim is to demonstrate the central role argumentation can play in the development of human-centric AI systems by providing computational models and tools for cognitive reasoning and dialogues among humans and machines at the global scale. We presented some preliminary ideas on this vision in Flouris et al. (2013) and Flouris et al. (2016); here, we elaborate more on these ideas, taking into consideration the recent advancements in related fields of research such as argumentation, machine learning and natural language processing.

The not-so-distant-future example that follows illustrates how we envision the interaction with the Web of Debates (Section 2). Section 3 gives more details about the vision: it motivates the need for its realization and describes how it will function, how people will benefit from it, and its main goals. Section 4 describes the challenges that stand in the way of its realization and proposes directions to overcome them, and Section 5 discusses its potential impact and some possible ethical issues that the Web of Debates may raise. Section 6 summarizes the main points of this vision paper.

2. Motivating example

The day began with a feeling of unrest for Steffi. The new article she is about to prepare obtains added gravity in the prospect of her country's elections next month. The topic is not unfamiliar to her; as a financial journalist she has written numerous articles in the past regarding the financial crisis and the impact of measures suggested by the International Monetary Fund (IMF) in other countries. Her intention this time is to question the diverse viewpoints on the IMF that are put forward by the different parties and to present as objectively as possible well-justified and clearly-articulated opinions both in favor and against the controversial role of IMF.

She hits "IMF policies help countries recover from financial crises" in ArgSE, the Arguments Search Engine she mostly uses when seeking for arguments on the Web, and configures its settings in "debate mode", in order to receive both supporting and refuting arguments. She has prepared a categorization of the different

target groups she is interested in to drive the mining process, and has uploaded the corresponding profiles using the “Audience Characteristics” functionality of ArgSE. For instance, she would like to know what arguments can be more meaningful for unemployed young people and middle-class workers.

Steffi has configured ArgSE to search for relevant arguments online but ignore sources with a low credibility score. Her profile data guides ArgSE to accurately decide on the level of detail to apply for the construction and presentation of arguments: her expertise in financial terms is sufficient to understand arguments on the connection between unemployment and inflation, but those regarding certain social aspects of unemployment require more detailed analysis in order to be comprehended.

As a result, ArgSE returns a graphic showing in a visually appealing manner the different arguments, as well as their relevant properties and metadata, including the sources (provenance) of each argument, the date and time of its publication, its supporting evidence, the argument style (e.g., deductive, inductive, etc.), its adequacy for a particular audience, and the relationships among the arguments (e.g., attack, support etc). It further identifies categorizations that Steffi did not consider in the first place, classifying certain arguments to audience groups sharing similar characteristics.

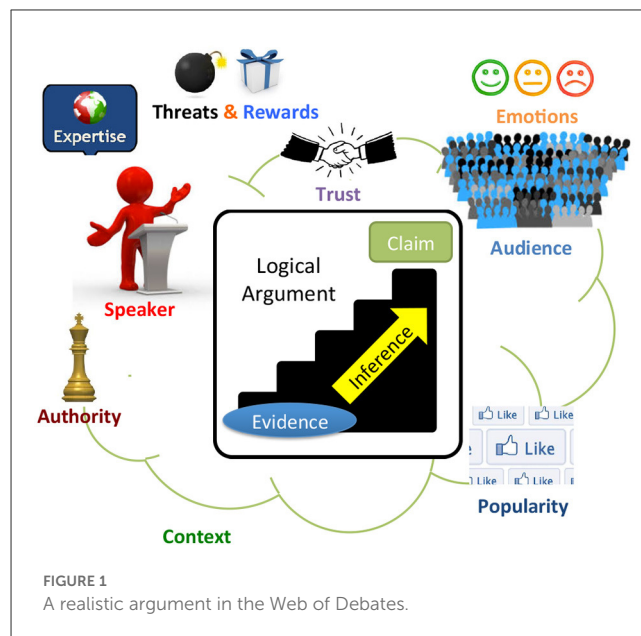
Using all the available information, Steffi navigates more deeply in the graph, she filters, questions, groups and organizes the available arguments, and eventually identifies and extracts the most convincing ones. A few hours later her article is ready. Her debate-enabled editor has assisted her in annotating the different parts of her text with a formal description of the arguments they refer to, so that search engines can identify and retrieve them, and links them with the respective online sources and evidence they are based upon. Steffi’s own conclusions, based on the correlation of facts she personally deduced during her research are also included (and annotated) in the text. This way, her annotated article and arguments can be stored in her electronic newspaper’s argument repository for others to find and reuse. As she sends the article to her editor she feels confident that her audience will have the means to form a well-informed opinion before actively participating in the country’s decision making process.

3. The vision of the Web of Debates

3.1. Why: the need for the Web of Debates

As the Web is increasingly being used for informational purposes, the public opinion is progressively being shaped by what people read online. Online versions of traditional mass media play a major role in this shift. On the other hand, due to the easiness with which content can now be uploaded, many users now use the Web as a podium to express themselves. However, extracting *meaning* out of the plethora of opinions (i.e., evaluating the credibility and coherence of information related to a subject of interest, understanding why it is important, and ultimately deciding whether to adopt or reject it) becomes increasingly difficult.

Even today’s Web contains the information necessary for Steffi to complete her article. However, this information, being in textual form, is not easily retrievable or processable, so it is not appropriate



for implementing the features presented in our example scenario. The Semantic Web (Berners-Lee et al., 2001) and Linked Data⁴ initiatives promised to overcome some of the limitations of natural-language Web pages by providing appropriate methodologies for publishing and interlinking semantic data on the Web using machine-processable formats. This has recently led to the development of *knowledge graphs* (graph-based representations of real-world knowledge; Hogan et al., 2021) and several types of knowledge-based systems, such as search engines, recommendation systems, personal agents, etc. However, the focus of these initiatives and models is on the representation of data, rather than arguments or opinions.

Similarly, the main tenets of computational argumentation (Besnard and Hunter, 2008; Baroni et al., 2018) and the extensive research conducted in this field have direct impact on the formulation of the new Web. This research has led to various types of applications in domains such as law, medicine, e-government and others (Atkinson et al., 2017). While they demonstrate well the potential of computational argumentation, they are all of small scale, being limited by their inability to process natural language arguments. On the other hand, the recent advancements in *argument mining* (Lawrence and Reed, 2019) have led to global-scale applications of argumentation such as argument search engines (Wachsmuth et al., 2017b; Stab et al., 2018; Chen et al., 2019). Their main functionality is to find on the Web arguments pro or con any controversial issue. *args.me* (Wachsmuth et al., 2017b) and *PerspectroScope* (Chen et al., 2019) rely on pre-structured arguments, while *ArgumentText* (Stab et al., 2018) has the ability to extract arguments from any Web document. They all rank arguments by relevance to the user-specified topic, while some of them present extra information for each argument such as supporting evidence, its stance score (denoting the extent to which it supports or refutes the claim) or its relevance score. While

⁴ <https://lod-cloud.net/>

these are closer to the kinds of applications we envision for the Web of Debates, their performance is still limited as, for example, evidenced by the results of a recent user-based evaluation, which showed that they do not significantly outperform conventional search engines especially with respect to the *convincingness* of the arguments they retrieve (Rach et al., 2020). This can be attributed on the one hand to the limitations of the argument mining methods they use, and on the other to the lack of a method to assess the quality or persuasiveness of arguments.

Realizing the types of services and applications we describe in Section 2 requires addressing the primal reason why opinions reach the Web in the first place, which is to be *persuasive*. This latter step is important, in order to depart from simple argument listings and logical argumentation, and support *realistic arguments* and *debates with a purpose*, i.e., debates where arguments are not-purely-logical, and have a certain aim, namely to persuade a certain audience on some topic, as happens in real-world debates, or help a group of people make an informed decision through deliberation.

3.2. How: the function and use of the Web of Debates

Current Web technologies focus on searching for and managing documents and information. The Web of Debates will additionally enable searching for and managing *realistic arguments* (Figure 1). A realistic argument will have an internal structure, containing a logical part, but also other types of information related to its persuasiveness or general quality: the audience that it is targeted at, its provenance, the context in which it was made, the values it promotes, the popularity of the claim that it supports, evidence for its believability (e.g., links to documents, facts, or other arguments that back it up), the conditions under which it is effective or valid, etc. Moreover, arguments will be *interlinked* in various ways, where the links may represent different types of support or attack relationships among the arguments (Figure 2). Understanding the role of the different components and interconnections of realistic arguments, as well as studying the factors that affect the persuasiveness and quality of arguments, such as emotions, trust, provenance, evidence and other logical or extra-logical considerations will be a crucial first step toward realizing the vision of the Web of Debates.

The Web of Debates will revolutionize the way argumentative information that exists on the Web is organized and exploited. Arguments will be uploaded directly by content providers, but it will also be possible to construct them on demand from text or by combining existing arguments with data from knowledge graphs and other types of knowledge bases, following formal methods, and taking into account the intended audience. To allow content consumers make the most out of the presented arguments, the Web of Debates will exploit information that is both of objective nature (e.g., the structure of an argument, the logical fallacies it may contain or its relationships with other arguments) and of subjective nature (e.g., the consumer's background knowledge and cognitive characteristics), based on which a proper ranking of the presented arguments will be possible, so that the strongest, most relevant and most understandable will be more visible. This, however, will not

undermine the *diversification* of the presented opinions. In order to prevent the formation of *echo chambers*⁵ or *filter bubbles*,⁶ the selection and ranking algorithms of the Web of Debates will ensure that arguments from all different viewpoints will be presented and highlighted, and the users will be allowed to access and configure the algorithms as they wish. In our motivating example, ArgSE would return the official opinions of IMF, as well as counter-opinions put forward by leading economists and other people, provided that they are trustworthy enough and understandable (per Steffi's knowledge background). It would also be able to explain how the presented arguments were selected and ranked and give the options to Steffi to configure the selection and ranking process.

Realistic arguments will be stored in "argument bases" (the analogous to knowledge bases and ontologies) and will be linked to online sources, such as a collection of sentences inside a document, information retrieved from a picture, etc. In the context of our example, people arguing about IMF's role in mitigating the effects of the economic crisis, will have the ability to post and interrelate arguments in a machine-interpretable way. Similarly, the IMF itself will be able to express its own arguments on the matter, stored in its own dedicated repository and uploaded on its website. Note that all types of digital artifacts (from financial reports to polls, simple text, images, videos, other arguments, datasets) can be used as evidence supporting a certain argument. Thus, arguments and digital objects will be interrelated in two ways: arguments can be linked to digital objects they refer to, whereas digital objects can also be used as parts of arguments (e.g., as supportive evidence).

The Web of Debates will also enable certain forms of *dialogical interaction* with its users. As described in the motivating example, after receiving a set of arguments that best match her request, Steffi will be able to follow up by requesting more arguments, by asking for more clarifying information about a certain argument, or even dispute the returned ones by presenting her own counter-arguments. ArgSE will then be able to search again in the Web of Debates and respond back by presenting additional persuasive arguments in the first case, data that back up or explain the argument in question in the second case, and data or arguments that invalidate Steffi's counter-arguments in the third case.

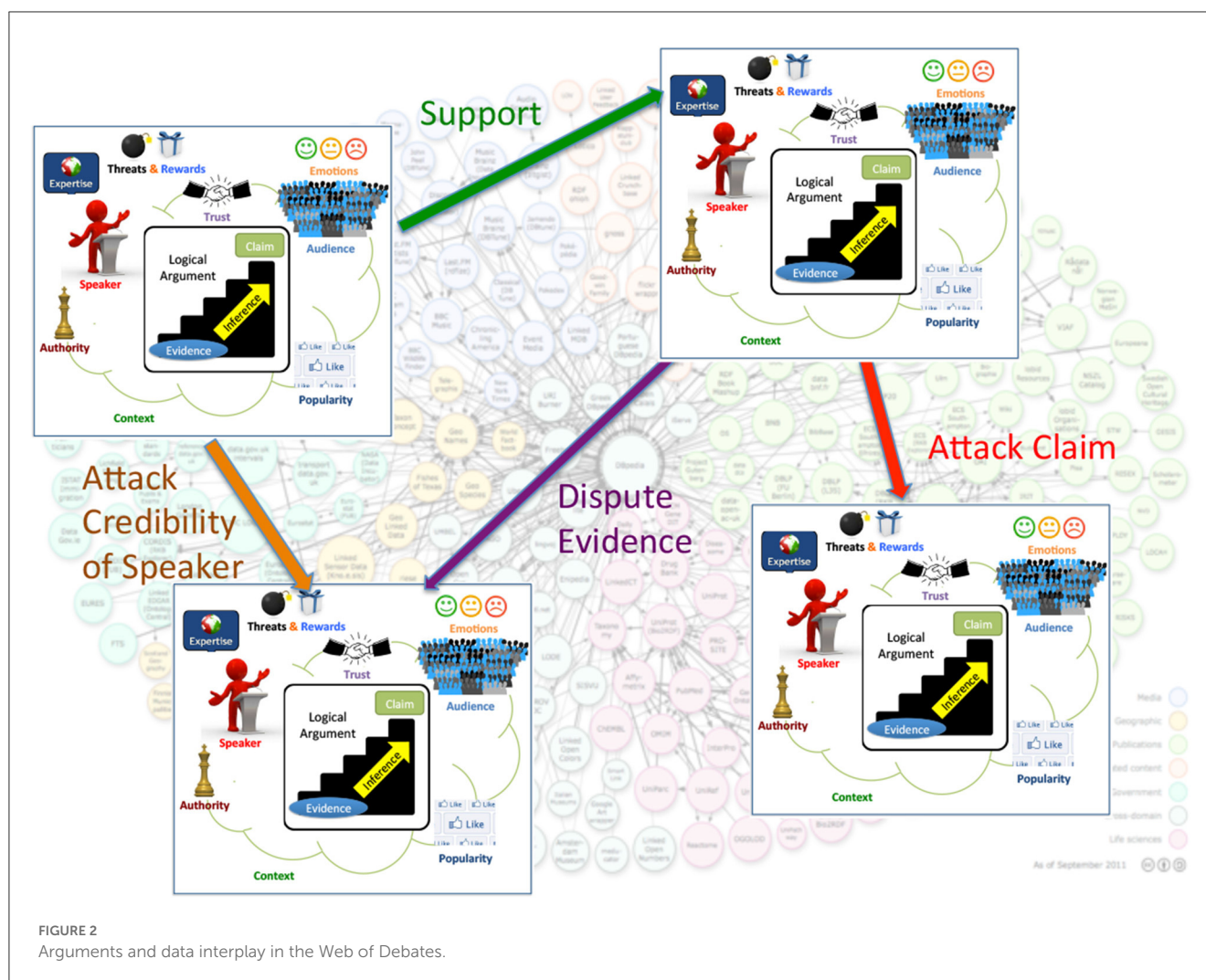
Summing up, we envision the Web of Debates not as a replacement for the current Web but as a complementary technology. Searching for and interlinking documents and information will still be among the core functions of the Web. The Web of Debates will provide additional tools that will exploit such functions to support a new one: the retrieval and management of arguments and the interlinking among arguments, web documents and information.

3.3. Who: actors in the Web of Debates

The Web of Debates will provide benefits for both the content provider and the content consumer, by offering a convenient podium for expressing one's opinions and a platform for accessing opinions of others. The easy access to the enormous amounts

⁵ https://en.wikipedia.org/wiki/Echo_chamber

⁶ https://en.wikipedia.org/wiki/Filter_bubble



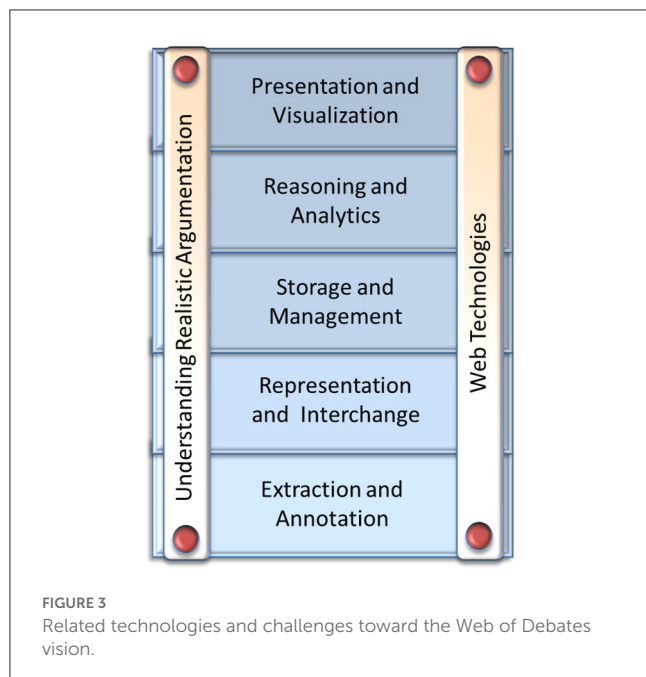
of Web information, in tandem with the automated annotation, retrieval, exploration and analysis of realistic arguments, will allow opinions to reach a large, literally global, audience, and, at the same time, provide a valuable tool in the hands of professionals, businesses, organizations, governments, or individuals to support their decision-making processes. This will be realized via the development of new and more powerful argument-aware search engines and other types of web applications that will allow users to retrieve, process, visualize, understand and query the arguments uploaded by content providers, as well as their interrelationships.

The combination of these features and tools will stimulate opinion diversity, contribute toward collective awareness and informed decision-making, promote active citizenship and e-democracy, support legal argumentation and justice attribution, allow improved fact-checking and encourage structured and civilized argument exchange in a networked world. In addition, it will help all parties formulate explicit opinions in their effort to persuade others into accepting a certain claim or taking a certain action, thereby using the Web to argue in favor of the products, services or ideas that they promote (for marketing or advertising purposes, or for refuting unjustified opinions or prejudices).

In our motivating scenario, Steffi is aided in her task by a graphic display summarizing the strongest arguments retrieved from credible sources on the Web, as well as their properties, supporting evidence and interrelationships. In this way, she would be protected from malicious users and sloppy arguments. Moreover, she would be able to concentrate on the most important ones or those that are most relevant to the specific context or case that she is interested in, and she would be able to easily identify poorly supported opinions.

3.4. What: the goal of the Web of Debates

The goal of the Web of Debates is not to impose any given opinion, but to provide the medium through which a user can “collect” different arguments in favor and/or against a certain claim in order to form an opinion of their own, convince an audience to accept a certain claim or opinion or participate in discussions with other users in order to take collective decisions about a certain course of action. The services offered by a search engine in the Web of Debates are analogous to those of a journalist, whose role is to



objectively and concisely reproduce the most prominent opinions expressed by different people or entities (e.g., political parties), in ways that help the readers better understand and evaluate them, taking into account their profiles and backgrounds. In our example, ArgSE retrieves and presents arguments from sources that are considered reliable, as well as information associated to their quality and persuasive strength for audiences that match the profiles provided by Steffi. But it is up to Steffi to decide which of them would actually be the most influential for the readers of the newspaper she is working for. Apart from search engines, the Web of Debates will support several other types of applications, such as everyday assistants, expert companion systems (see e.g., Dietz et al., 2022 for some examples), collaborative decision support systems, intelligent tutoring systems aimed at teaching users how to make better arguments, automated debating systems and others. Some common characteristics of all such systems will be their focus on natural-language arguments and the human aspects of argumentation, their seamless integration within the online, private or social, activities of their users, their adaptability to background knowledge and cognitive characteristics of each user or group of users, their ability to explain any inferences they make, and their ability to develop by learning from experience and by taking into account the feedback provided by their users. In other words, they will combine all major characteristics of *human-centric* AI systems.

4. Realizing the vision

There are several research fields and state-of-the-art technologies that can provide the substrate upon which the vision of the Web of Debates can be realized, but also important obstacles that stand in the way of its realization. Figures 3, 4 provide an overview, showing some broad research fields and technologies that are relevant.

Figure 3 lists the main relevant technologies. The vertical bars represent various challenges that need to be overcome by the corresponding technologies and research fields. The horizontal bars represent critical technologies, which, even though not directly used to address any challenge, will set the guiding principles upon which the solutions to all challenges will be based. All these technologies need to be advanced or further explored to overcome the related challenges.

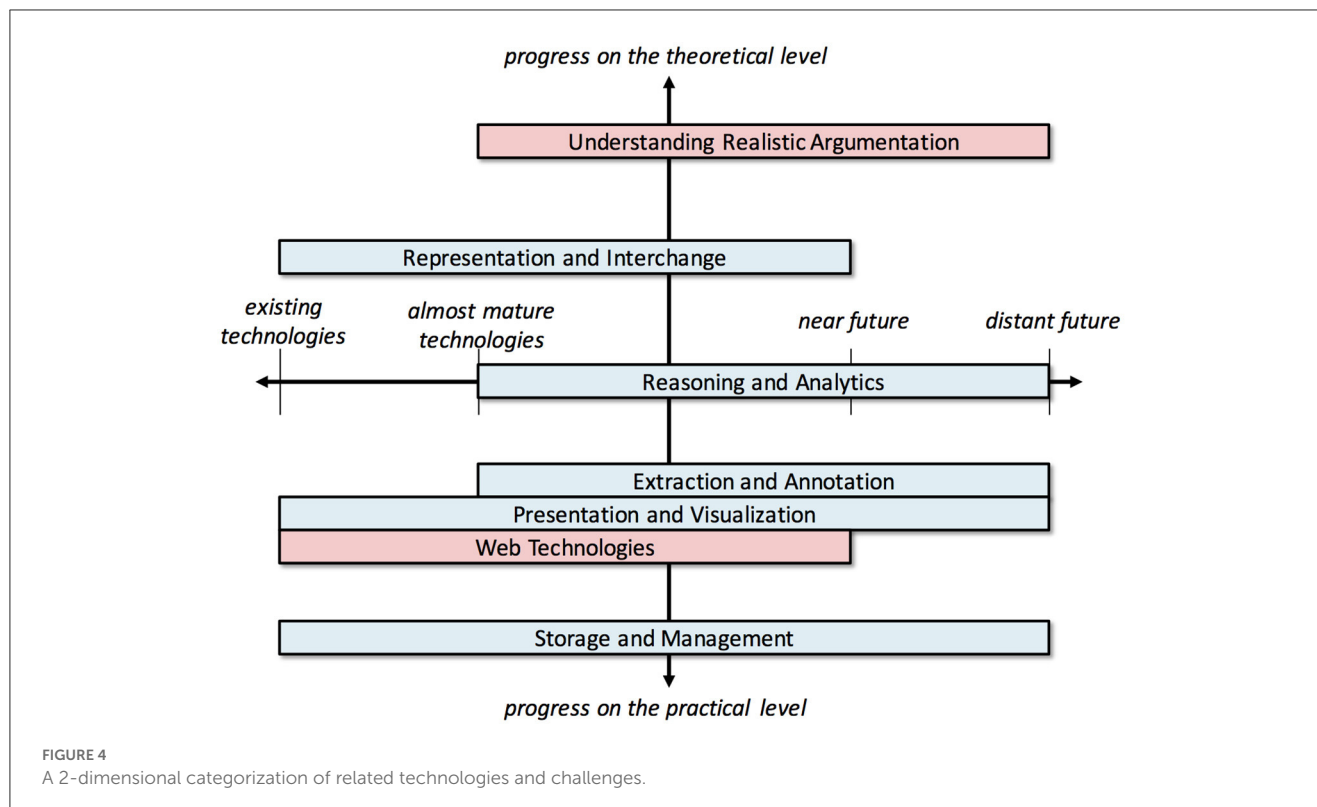
Figure 4 displays the same technologies and challenges from a different perspective, organizing them in a two-axis chart. The position of the technology along the horizontal axis represents both the current and the required maturity of each technology to solve the respective challenge. The left side of each rectangle represents the current capacity of the corresponding technology to address the related challenge, at least at a preliminary stage, whereas its right side represents additional advances that need to be achieved (and how far in the future these are estimated to occur) before actually solving the respective challenge in its entirety. On the other hand, the vertical axis represents the kind of progress required per technology (practical or theoretical) to overcome the respective challenge. We should note that this chart is based on our own assessment of the maturity level of each technology based on the literature we reviewed, and not on a systematic evaluation of the technologies.

In the following sections, we further analyse these technologies and their role in the realization of the Web of Debates.

4.1. Understanding realistic argumentation

Argumentation theory studies how conclusions can be reached through logical reasoning in the presence of, possibly contradictory, evidence for or against a certain conclusion, whereas *argumentation systems* are logic-based computational systems that aim to automate this process (see Baroni et al., 2018; Gabbay et al., 2021 for the state of the art and current trends). Scientific advances in these fields of study, such as the understanding of the structure of arguments, the development of tools for constructing arguments, the identification of their relations, and the development of semantics for drawing sound logical conclusions from possibly contradictory arguments, are all relevant in the context of the Web of Debates.

Nevertheless, the Web of Debates is a lot more than an argumentation system deployed in a global scale. The main challenge here is the shift from *logical argumentation* to *realistic argumentation*. Realistic argumentation does not only appeal to the logic of the audience, but also to its emotions. It is only partly based on facts and data, often employing additional techniques such as the clever use of verbal cues and the semantic structure of text/speech (politeness, aggressiveness etc), as well as different *argument schemes* based on factors such as appeal to authority or expert opinion, popularity of supported claims, peer-pressure, arguments from analogy, proof-by-example, non-logical (e.g., statistical) correlations between different arguments, and others (Walton, 2006). The aim of realistic argumentation is usually to *persuade* or *help reach a decision*, rather than prove or present facts or arguments for the sake of presenting them; thus, it also



involves a process of selecting the strongest arguments to put forward first, taking into account their *relatedness*, *informativeness* or *persuasive* characteristics. In this sense, realistic argumentation is more context-aware and more personalized.

Building on the most influential model of arguments in the last decades, *Abstract Argumentation Frameworks* (Dung, 1995), there have been some attempts to formalize features of realistic argumentation, such as the audience (Hunter, 2015), the values that arguments promote (Bench-Capon, 2003), preferences (Amgoud and Vesic, 2014), trust (Villata et al., 2013), the argument strength (Amgoud et al., 2022), the context of argumentation (Brewka and Eiter, 2009), uncertain arguments (Hunter, 2013), commonsense arguments (Vassiliades et al., 2020), enthymemes (Black and Hunter, 2008), and persuasion dialogues (Prakken, 2009). There is also some promising research on the formalization of argumentation schemes (Verheij, 2003; Reed and Walton, 2005; Prakken et al., 2015; Wyner, 2016; Panisson et al., 2021), and more generally on the use of argumentation schemes in AI (Macagno, 2021). The study of what contributes to the persuasiveness or the quality an argument has recently started but is growing fast. Work on this topic includes crowdsourcing studies comparing arguments in terms of their persuasiveness (Habernal and Gurevych, 2016; Gretz et al., 2020), studies focusing on specific factors such as linguistic features (Persing and Ng, 2017), the semantic types (logos, ethos, or pathos) of claims and premises (Hidey et al., 2017), the types of evidence used to support an argument (Addawood and Bashir, 2016), personality traits and prior beliefs of the audience (Lukin et al., 2017; Durmus and Cardie, 2018; Al-Khatib et al., 2020), the

style of the arguments (Baff et al., 2020), etc., but also some more general attempts to identify all related factors (Steenbergen et al., 2003; Wachsmuth et al., 2017a). According to Wachsmuth et al. (2017a), the quality of an argument is determined by its *cogency*, i.e., whether its premises are acceptable, relevant to its conclusion and sufficient to draw its conclusion; its *effectiveness*, which is related to the credibility of its author, its clarity, its emotional appeal and its appropriateness; and its *reasonableness*, which refers to its global acceptability, its relevance to the discussion or debate, and its ability to defend itself against all counter-arguments. The deliberative quality of an argument, defined in Steenbergen et al. (2003), includes additional factors that are important in deliberation dialogues, such as *respect*, *equality* among all arguers, *interactivity* and *testimoniality*. Most work in this area has the form of empirical studies aiming at validating the related factors, and improving our understanding of human argumentation. There are still, though, relevant issues from the perspective of discourse analysis, rhetorics, and psychology [e.g., whether people are skilled arguers (Hahn and Oaksford, 2012), and why people argue (Mercier and Sperber, 2011)] that has not yet attracted much attention from the AI community. Some other open research problems in this area concern the interaction of the different factors, how teams of arguments work in concert in debates, how the order that the arguments are presented influence the outcome of a debate, and how people select which arguments to put forward in a debate. The interdisciplinary study of such issues is necessary for understanding and formalizing human argumentation, which is in turn a key requirement for realizing the Web of Debates.

4.2. Web technologies

The current Web is based on the simple idea of interlinking documents and making them available to anyone from anywhere. Building on the same principle, different technologies have been proposed to extend the document Web. One of the most prominent ones is the *Semantic Web* (Berners-Lee et al., 2001) and the closely related *Linked Data* initiative, where the main building blocks are structured datasets (rather than documents). Its motivation is that documents are not easily machine-processable, so there are certain limitations on what a machine can do with them; on the other hand, access to machine-interpretable data (in the sense of a “global database”) can give rise to even more sophisticated applications, such as the ones that have already been created on top of knowledge graphs (see Hogan et al., 2021 for some examples).

The so-called *Social Web* aims to foster social interaction, by providing a plethora of tools and platforms enabling humans to communicate through blogging, tagging, Web content voting, social bookmarking, and other means of social interaction. The Web of Debates seeks to upgrade the role of the Social Web into a broader means of communicating opinions and carrying out debates. There have already been some attempts to integrate argumentation within the Social Web. For example, Schneider et al. (2013) provides a review of web applications that combine features of the Social Web, the Semantic Web and computational argumentation. Such applications, however, are still limited in the features of realistic argumentation they can support as they mostly rely on models that capture the logical aspects of argumentation. Frameworks for social argumentation (Leite and Martins, 2011; Baroni et al., 2015; Patkos et al., 2016) integrate arguments with social votes; online debates, though, involve a lot more non-logical aspects, which these frameworks do not capture. With a shift toward realistic arguments, knowledge exchange will be carried out along the lines of logical consistency, factual accuracy and some degree of emotional appeal to the intended audience, but will also take into account the individual needs and preferences of web users. Even though the decision of adopting one conclusion over another will remain a subjective issue, the Web of Debates will facilitate the process of deliberation by filtering out irrational and logically incorrect expressions, while maintaining a significant degree of personalization in choosing the top-rated arguments for each user.

The *Pragmatic Web* (Schoop et al., 2006) is motivated by the observation that the content of the Web does not actually represent factual data, but the subjective opinions of the people who upload it. Even though it has a similar motivation with the Web of Debates, its objectives and used methodologies are quite different. From the Pragmatic Web viewpoint, a conflict is just a clash of opinions, which is resolved not by analyzing the opinions themselves, but by determining the support of each opinion via crowdsourcing techniques, and by interpreting and representing data in a context-dependent manner so as to enable users to reach agreements. On the other hand, the Web of Debates aims to analyse and contrast the different contradicting arguments, to allow the interested user to better understand their connections, and eventually judge themselves the validity of each one, based on their own beliefs, knowledge, or even prejudice; unlike the approach followed by Pragmatic Web, this would allow the identification of widely spread, but unjustified, beliefs or opinions.

Closer to our vision is the *Argument Web* (Bex et al., 2013; Reed et al., 2017), which is an effort to deploy argumentation on the Web. At its core is the *Argument Interchange Format* (AIF, Chesñevar et al., 2006; Rahwan et al., 2007), an ontology for arguments. On top of AIF, several Web-based tools have been developed for argument annotation, visualization and analysis⁷ and have been applied to various types of real debates, including, for example, debates taking place in the famous BBC broadcast Moral Maze.⁸ Other applications include tools for better understanding existing arguments, or for improving the argumentation skills of adolescents.⁹ All these developments are in line with our vision of the Web of Debates and will contribute to its realization. These tools, however, rely mostly on manual annotation and analysis and cannot, therefore, meet the requirements of large-scale applications. The realization of the Web of Debates will require the automation of the argument annotation and analysis processes, their enhancement so that they can handle all features of human argumentation, and the development of several other extra-logical processes, such as profile and context analysis, audience analysis, trust analysis, reputation analysis and others. This will enable the development of large-scale web applications that can take advantage of all argumentative information that is already available on the Web.

In summary, the technological advances made in the context of the above technologies will contribute to the development of the Web of Debates in a critical manner. In particular, the low-level infrastructure of the Web of Debates is expected to reuse the standard Web protocols, whereas knowledge graph languages and semantic technologies, and other techniques and technologies such as crowdsourcing, social tagging, voting and others, which Web users are already familiar with, will probably find their way into the Web of Debates. The developments made in the Argument Web with respect to argument modeling, annotation and visualization will also be exploited and extended or adapted to the needs of the Web of Debates.

4.3. Extraction and annotation

As with all added-value technologies, the size of the Web of Debates must reach a critical mass to make itself useful. Given the abundance of the natural language arguments already on the Web, technologies such as automated mining of arguments from blogs, forums or other social media, Natural Language Processing (NLP) techniques and others, need to be employed to create structured arguments out of text. In addition, human contribution could be enabled for this task, by adapting existing technologies such as *gamification* (von Ahn and Dabbish, 2008) or *crowdsourcing* techniques. Some efforts have already been made to crowdsource argument creation (Chalaguine and Hunter, 2019) and annotation (Ghosh et al., 2014; Skeppstedt et al., 2018). Furthermore, aspects

⁷ <https://arg-tech.org/index.php/research>

⁸ <https://www.newsweek.com/artificial-intelligence-argument-debate-752199>

⁹ <https://www.independent.co.uk/tech/artificial-intelligence-debate-argue-bbc-science-tech-research-a8118191.html>

related to *multilinguality* should be addressed, exploiting the improving quality of automated translation tools. Along similar lines, the annotation of images, sounds or complete documents with the arguments that characterize them is equally critical for a Web where knowledge can take various forms.

In tandem with the above efforts, it is of crucial importance to encourage content providers to upload their arguments online using the proper format (i.e., in a structured form), by providing tools that simplify the process, e.g., by allowing the semi-automatic generation of arguments and/or by aiding the content provider to annotate her arguments. Existing tools for manual argument creation or annotation, such as Araucaria (Reed and Rowe, 2004), Rationale (van Gelder, 2007), OVA (Bex et al., 2013), and Carneades (Gordon et al., 2007), enable the users to identify the components of arguments (e.g. their premises, conclusions, etc.), their relations (e.g., attack, support, etc.) and the argumentation schemes they instantiate (e.g., argument from expert opinion, etc.).

However, in order to be able to exploit the abundance of natural language arguments that already exist on the Web, automating the extraction of arguments from text is a fundamental requirement. The rapidly expanding field of *argument mining* (see Stede and Schneider, 2018; Lawrence and Reed, 2019 for a recent survey and book) has already demonstrated some promising results that could form the basis for realistic argument extraction and annotation in the Web of Debates. These include annotation schemes for argument mining (Budzynska and Reed, 2011; Peldszus and Stede, 2013; Stab and Gurevych, 2014; Kirschner et al., 2015; Habernal and Gurevych, 2017; Niculae et al., 2017), annotated corpora (Andreas et al., 2012; Ghosh et al., 2014; Rosenthal and McKeown, 2015; Abbott et al., 2016; Habernal and Gurevych, 2017), methods for argument extraction from text (Andreas et al., 2012; Florou et al., 2013; Ghosh et al., 2014; Rosenthal and McKeown, 2015; Abbott et al., 2016; Habernal and Gurevych, 2017) or for identification of argument relations (Peldszus and Stede, 2015; Cocarascu and Toni, 2017; Lawrence and Reed, 2017; Niculae et al., 2017; Nguyen and Litman, 2018; Kobbe et al., 2019; Trautmann et al., 2020). Most of the current corpora and argument mining methods have been developed for specific domains and applications and the performance varies across different tasks and domains; for example, the results are much better in persuasive essays (Stab and Gurevych, 2017) than in legal cases (Teruel et al., 2018) or microtexts (Peldszus and Stede, 2015), which are most commonly encountered on the Web. There is still lack of a general annotation scheme and generic methodologies that would perform well in multiple domains. We should note here that it may be impossible to develop a computational method that can with 100% accuracy identify arguments in a natural language text. As evidenced by several studies that involved manual annotation of texts (Stab and Gurevych, 2014; Kirschner et al., 2015; Habernal and Gurevych, 2017), there is very often disagreement between annotators on the arguments, components of arguments or argument relations conveyed by a text, which in most cases is due to the ambiguity of human language. As shown in Thorn Jakobsen et al. (2022), it may also be due to the different backgrounds and demographic characteristics of the annotators. Manual annotation may therefore introduce *social bias* to the data used to train data-driven argument mining methods and, as a result, also to the methods themselves.

Addressing this challenge is a requirement for the realization of the Web of Debates, while methods for identifying and measuring biases (Pagano et al., 2023) can also help mitigate this issue. Most current argument mining approaches focus on arguments, components of arguments (e.g., premises and claim) or relations between arguments (e.g., attack and support). There have been some attempts to automatically extract from text other features of human argumentation such as ethotic expressions (Duthie and Budzynska, 2018), emotional arguments (Oraby et al., 2015) and argument schemes (Lawrence and Reed, 2016), but the research in this area is still in its early stages. Developing domain-independent methods with the capability of identifying extra-logical features of argumentation is essential for the development of solutions that better fit the needs of the Web of Debates.

4.4. Representation and interchange

Enabling the association and combination of arguments from different sites of the Web requires the development of a semantically explicit representation model (ontology) for realistic arguments, so that different independently developed applications will be able to process them in a common manner and interoperate within an integrated environment. As also discussed above, AIF (Chesñevar et al., 2006; Rahwan et al., 2007) is one such ontology, which captures various models of argument, both formal (such as AAFs), and informal such as Walton's *argumentation schemes* (Walton, 2006). Using AIF, it is possible to model the (logical) structure of an argument (e.g., its premises, conclusion, etc.), argument relations (e.g., support, conflict, preferences), but also the argumentation scheme that an argument adheres to. An extension of AIF enables also modeling elements of argumentative dialogues such as locutions (e.g., statements, withdrawals, questions, challenges, etc.), commitments and dialogue rules (Reed et al., 2008). Such approaches are definitely within the spirit of the Web of Debates. There are still though several aspects of human argumentation that have not been accommodated. The development of an appropriate model for realistic arguments requires answering additional questions such as: What are exactly the types of information that define the quality or persuasiveness of an argument? How are these modeled and attached to an argument? How do we characterize and model the presenter of an argument and her audience? What are the possible relations between realistic arguments and the possible statuses of an argument within a realistic debate? Most of these issues are still open research topics in computational argumentation, with some interesting approaches being proposed during the last few years (e.g., see Bench-Capon, 2012).

The representational model will be based on knowledge graph languages, to allow reusing existing ontologies that capture features related to realistic argumentation [e.g., profile ontologies such as UPOS (Sutterer et al., 2008) or provenance ontologies such as PROV-O (McGuinness et al., 2013)], and exploiting the Linked Open Data (LOD) architecture to provide connections between the concepts/topics related to the arguments and their representation in existing online datasets (e.g., Wikidata). This will enable interlinking related arguments, but also linking arguments

with other types of web data, which can be used for example as supporting evidence. It will also allow using standard Semantic Web languages and tools (e.g., SPARQL, rule languages, etc.) for querying and reasoning with the arguments and their relationships.

4.5. Storage and management

Realistic arguments will be stored in what we call “*argument bases*”, the analogous of knowledge bases. Their structure will enable storing arguments, as well as any other information that is relevant to the proper representation of realistic arguments and debates. Argument bases should also provide: (a) inference support; (b) query support; (c) support for data management tasks such as updating, repairing and change monitoring; (d) alignment and interoperating capabilities with related ontologies; and (e) propagation of relevant information among different systems. For the development of such systems, the experience gained from the deployment of triple stores and other semantic data management systems (Özsu, 2016; Abdelaziz et al., 2017) will be exploited. The AIFdb database system (Lawrence et al., 2012), which was developed for storing and managing arguments described in the AIF ontology, supports some of the desired functionalities: it enables semantic processing and visualization of arguments, query management and dialogue control. A language for querying structured dialogical data, which is compatible with AIF and knowledge graph languages (RDF, SPARQL), was also recently developed (Zografistou et al., 2018). Such technologies are compatible with and can form the basis for the development of web-scale argument bases for the Web of Debates.

4.6. Reasoning and analytics

Representing and storing arguments in an adequate format is not an objective in itself, just the means toward providing adequate services over the Web of Debates, based on the general notions of analytics and reasoning. Through these services, the user will be able to search and navigate through arguments (possibly in an exploratory manner), pose structured queries over the pool of available arguments, or perform sophisticated (and customized) aggregation and summarization operations. In addition, sophisticated forms of reasoning may emerge, allowing the identification of implicit relationships among arguments, or the development of new forms of semantics that determine the “acceptability” of realistic arguments, along the tradition of abstract argumentation (Dung, 1995). There are already several tools, called *argumentation solvers*, that were designed to solve standard reasoning tasks (e.g., compute the set of acceptable arguments) in abstract argumentation frameworks—see Cerutti et al. (2017) for an overview and Lagniez et al. (2021) for the results of the latest International Competition on Computational Models of Argumentation. The standard acceptability semantics of AAFs, proposed in Dung (1995) and considered in all these tools, use two (*accepted/rejected*) or three values (*accepted/rejected/undetermined*) for representing the acceptability of arguments. This is, however, too simplistic

compared to the way that we evaluate arguments in our every day life, where we most commonly believe in or are persuaded by arguments to varying degrees. This has recently led to finer-grain gradual evaluation methods, based on numerical scales (Baroni et al., 2019) or rankings (Bonzon et al., 2016). Some of these approaches also consider a *base weight*, a value assigned to an argument, which may represent the probability of believing the argument (Hunter, 2013), the aggregated strength of its premises and inference rules (Spaans, 2021), votes provided by users (Leite and Martins, 2011), the importance degree of a value promoted by the argument (Bench-Capon, 2003), or the trustworthiness of the argument’s source (da Costa Pereira et al., 2011). Extending these methods to take into account the factors associated with the persuasiveness or quality of arguments discussed in Steenbergen et al. (2003) and Wachsmuth et al. (2017a) (see also Section 4.1) is a promising research direction that would contribute to the realization of the Web of Debates. A computational framework that combines an arbitrary set of factors to compute the overall quality or acceptance of an argument was proposed in Patkos et al. (2016); however, the framework is generic and takes only into account the users’ arguments and votes. Further research is required to determine the extent to which each factor contributes to the quality of an argument, possible dependencies among the factors, and the role of the topic or context of a debate in determining which factors are more or less important.

Another aspect that should be taken into account is the much bigger scale of the Web of Debates compared to current argument-based applications. The majority of the reasoning problems in AAFs are known to be NP-hard (Charwat et al., 2015), and reasoning with realistic arguments is expected to be even more complex. The exact and complete solutions implemented by argumentation solvers may not, therefore, be feasible in scenarios involving large scale datasets. There have already been some recent efforts to develop approximate solutions for AAFs based on graph neural networks (Kuhlmann and Thimm, 2019; Craandijk and Bex, 2020; Malmqvist et al., 2020). The realization of the Web of Debates will require the development of similar approximate solutions for the evaluation of realistic arguments.

The *automated generation of arguments* on the basis of data or other arguments found on the Web will also be a desirable feature for many applications of the Web of Debates. This will create additional value from existing arguments, via aggregation, summarization, elaboration, and generation of new knowledge in the form of new realistic arguments. This is similar to how reasoning and inference generates new knowledge from existing facts based on well-defined formal deductive rules. In this direction, the approach proposed in Khatib et al. (2021), where arguments are generated by GPT-2, a neural language model, trained with data from argument knowledge graphs, has demonstrated promising results and a methodology that fits the envisioned features of the Web of Debates.

4.7. Presentation and visualization

Given the sheer size of the Web, one expects to find a large number of arguments in favor (or against) a certain claim, so

presenting everything to the user is certainly not productive. Some kind of *aggregation* or *summarization* is necessary, along with a *ranking* process that will highlight the most important or relevant ones, taking into account also issues like the diversification of opinions. It should be emphasized that ranking only aims at the practical necessity to give priority to some of the arguments; the user should have access to all arguments, and no filtering or censorship should take place as part of the ranking process. Preliminary research in this area has focused on identifying similar arguments using clustering techniques (Misra et al., 2015; Boltuzic and Snajder, 2016) and on summarizing the key issues brought up in debates using standard text summarization techniques (Ranade et al., 2013), tools and techniques from lexical semantics (Saint-Dizier, 2018), or machine learning techniques and word embeddings (Misra et al., 2017).

A similar challenge is related to the *visualization* of arguments and their relationships, which is important for the content consumer to understand the structure of a complex web of realistic arguments. Tools such as Araucaria, Rationale, OVA, and Carneades (discussed in a previous section) visualize debates as trees or graphs, focusing on the logical part of arguments or their relationships. Other argument mapping tools are Kialo,¹⁰ which displays one argument at a time with its support arguments on one side and the attacking arguments on the other, and DebateGraph,¹¹ which also focuses on one argument at a time and displays its related arguments in the form of a graph. Some of these tools display additional data about the arguments, such as a score or links to related debates or data. Such data but also any other information that is related to the quality or persuasiveness of an argument should be somehow made available to the users of the Web of the Debates and visualized in an intuitive way that will help them make sense of all different viewpoints in a debate as quickly as possible. Addressing the tradeoff between making available all relevant information to the users while, at the same time, helping them to make sense of a debate as quickly as possible is definitely a big challenge, and will require the adoption of standard information visualization principles such as the ones proposed by Shneiderman (1996), i.e., *overview, zoom and filter, details on demand, relate, history and export*.

5. Impact of the Web of Debates

5.1. Potential impact

The Web of Debates can be viewed as the “blog of tomorrow”, where people will be able not only to express their viewpoints in a natural language, but also to annotate and connect them in a machine-interpretable way. The expression of arguments in formal, machine-processable terms, as well as their interlinking, will create significant added-value benefits. In the same way that linked data and knowledge graphs have led to the discovery of new, previously unseen connections, correlations and knowledge (e.g., business analytics), we expect the interlinking of arguments to lead to a

better understanding of the various debates and the generation of new, aggregated or previously unknown arguments and insights.

The abundance of Web data, combined with machine-processable arguments, will allow the envisioned version of the Web not only to provide relevant information (as when reading a book), but also to combine available data in order to provide arguments in favor of (or against) different alternative options (as done by a knowledgeable expert). This way, people will be better informed on matters of interest, thus promoting collective awareness on community problems and enabling better decision-making for professionals or companies.

At the community level, the services of the Web of Debates can enable public authorities to reach a broader audience in a more personalized way, in order to foster policies of societal value (e.g., healthy lifestyle, sound environmental behavior), to target unjustified concerns, to promote participation in community matters and democratic processes (e-democracy), or to support legal argumentation and justice attribution. At the individual level, the same services are expected to form a critical component of future autonomous entities endowed with socio-cognitive intelligence, which are used in the emerging market of smart spaces (Alazab et al., 2022). This can find applications ranging from service robots for domestic use, to smart environments related to domestic care and work, education, healthcare, communication and entertainment.

In addition, there is a wide range of potential applications suitable for the private sector; these generally fall under marketing, e.g., persuading customers to buy products/services, convincing people to donate to a charity, etc. Similarly, the Web of Debates can also be used as an assistive tool for individuals that practice persuasion as part of their professional life, such as lawyers, business executives etc, or for decision-makers in general, as it would allow better and more informed choices by combining information found on the Web, and also possibly in local databases, to build persuasive arguments and suggestions. But at the same time, by relying on transparent and easily configurable algorithms that promote the diversification of the viewpoints they present to the users, it can also help mitigate the problem of echo chambers and the increased polarization that this phenomenon causes.

Ultimately, we see the Web of Debates as the platform of ideas that holds the promise for promoting the role of humans in collective decision-making and e-democracy, able to have significant impact at both the individual and the societal level.

5.2. Ethical issues

The ability of the Web of Debates to adapt to the personal characteristics and background knowledge of each user requires that it has access to this information. However, it is important to ensure both that the users will be in total control of their personal data, and that the functionality of the Web of Debates will not be diminished by the lack of personal data. This can be ensured by developing the Web of Debates according to the *Privacy by Design* principles (Cavoukian, 2013). Following these principles, the Web of Debates should by default not have access to any personal data, its operations should be visible and transparent to

¹⁰ <https://www.kialo.com/>

¹¹ <https://debategraph.org/>

all users, it should provide several data-sharing options that will be easily comprehensible to all users, and it should employ end-to-end security mechanisms for protecting the users' data.

We acknowledge the fact that persuasion (that underlies the Web of Debates), as well as the development of automated persuasion systems, would, by their very nature, be open for misuse by governments, businesses, individuals or organizations (e.g., for coercion, control or opinion enforcement). For example, one potential issue would be the usage of the Web of Debates as a means to promote the incorporation of false, deceptive or misleading arguments by malicious content providers. In both cases, naive content consumers could be deceived, thus causing disillusionment to well-intentioned users and jeopardizing the usefulness of the Web of Debates.

Despite the fact that such opportunities for abuse are admittedly present, this is the case for most useful technologies, so we argue that this should not be a deterring factor toward realizing this technology. As a most striking example, one could refer to today's Web, where all such features exist (inaccurate or false information, etc.). However, we argue that the Web of Debates will in fact improve the situation, and will be helpful toward mitigating this problem.

In particular, it should be noted that it is not the aim of the Web of Debates to provide any kind of censorship or checking on different opinions. On the contrary, it will allow all opinions to be more easily publishable and accessible. We argue that this feature will in fact reduce the opportunities for censorship, coercion, or deception, in the sense that access to different opinions, as well as the verification of the validity of arguments associated with these opinions, will be easier for open-minded content consumers, so the power of deceptive or misleading arguments and opinions will be mitigated.

Similarly, understanding persuasion (in general) can reduce the opportunities of coercion, control, or manipulation that may potentially be exercised by businesses, individuals or organizations over unaware citizens. Research on persuasion can help in identifying how and when this happens, as well as in preventing it, by allowing humans and intelligent systems to argue together.

At a more technical level, advances in the fields of trust and automated fact-checking,¹² as well as the incorporation of provenance information in realistic arguments could help users in the task of identifying deceptive or misleading arguments. This is similar to how the current Web has allowed recent advances in technology where facts and statements can be more easily checked for validity against the vast amount of the information available on the Web, using fact-checkers.¹³

Furthermore, the integration of models and methods from Explainable Artificial Intelligence (Banerjee and Barnwal, 2023), especially in the processes that involve Machine Learning algorithms (e.g., argument mining or argument generation) will contribute to the transparency, interpretability and understandability of the outputs of the Web of Debates tools and applications and to the establishment of trust with their users. Computational argumentation has already proved to be a very

useful tool for developing explainable systems (Vassiliades et al., 2021), while the recent launch of the International Workshop on Argumentation for Explainable AI¹⁴ shows that this is an active area of interest for researchers in computational argumentation. We, therefore, anticipate that their involvement in the design and development of the Web of Debates will ensure that it will function as an explainable system.

6. Conclusion

Not long ago, the problem of information overload attracted the attention of different scientific communities, fueled by the increasing number of people posting and accessing information on the Web; nowadays, the increasing amount of user-generated reviews, comments and arguments on the Web may lead to a similar problem, that of opinion overload. In this paper, we looked ahead to a future version of the Web, where this problem can be overcome by exploiting the structure of realistic arguments and understanding the arguers' intentions. After motivating and describing our vision, we identified its main challenges and proposed research and technological directions to its realization, which can be summarized in: understanding and formalizing realistic arguments and debates; developing methods and tools for automatically generating structured arguments (e.g., by extracting arguments from text); developing appropriate models for the representation and interchange of arguments; creating systems for their storage and management; developing methods for analyzing arguments and debates; developing models and methods for summarizing and visualizing arguments and debates; and augmenting Web technologies with the ability to automatically process online arguments by integrating the above research developments.

We strongly believe that the realization of this vision will stipulate research in a wide range of domains—scientific, academic and commercial—and can lead to the development of innovative human-centered applications that will revolutionize Web experience. Apart from its evident impact on the organization of argument and knowledge exchange on the Web, this effort opens up a way to serve a higher-level purpose: by enabling people to locate the valid rational arguments in the sea of opinions of questionable credibility, as well as those arguments that better support them, it will empower critical thinking and facilitate the active participation of humans in collective governance processes. Ultimately, we see the Web of Debates as the platform of ideas that holds the promise for promoting the role of humans in collective decision-making and e-democracy, able to have significant impact at both the individual and the societal level.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material,

¹² http://en.wikipedia.org/wiki/Fact_checking

¹³ https://en.wikipedia.org/wiki/List_of_fact-checking_websites

¹⁴ <https://people.cs.umu.se/tkampik/argxai/2022.html>

further inquiries can be directed to the corresponding author.

Author contributions

AB, GF, TP, and DP contributed to the conception of the main ideas. AB, GF, and TP contributed to the review of the relevant literature. AB and GF contributed to the revision of the paper. All authors contributed to the first version of the paper and read and approved the submitted version.

Funding

GF, TP, and DP were partially supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the 1st Call for H.F.R.I. Research Projects to support Faculty Members and Researchers and the procurement of high-cost research equipment Grant (Project Number: 4195).

References

- Abbott, R., Ecker, B., Anand, P., and Walker, M. A. (2016). "Internet Argument Corpus 2.0: an SQL schema for Dialogic Social Media and the Corpora to go with it," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*, eds N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odiijk, and S. Piperidis (Portorož: European Language Resources Association).
- Abdelaziz, I., Harbi, R., Khayat, Z., and Kalnis, P. (2017). A survey and experimental comparison of distributed SPARQL engines for very large RDF data. *Proc. VLDB Endow.* 10, 2049–2060. doi: 10.14778/3151106.3151109
- Addawood, A. and Bashir, M. N. (2016). "What is your evidence? A study of controversial topics on social media," in *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016* (Berlin: The Association for Computer Linguistics).
- Alazab, M., Gupta, M., and Ahmed, S. (eds.). (2022). *AIoT Technologies and Applications for Smart Environments*. London: The Institution of Engineering and Technology.
- Al-Khatib, K., Völke, M., Syed, S., Kolyada, N., and Stein, B. (2020). "Exploiting personal characteristics of debaters for predicting persuasiveness," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, eds D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetrault (Association for Computational Linguistics), 7067–7072.
- Amgoud, L., Doder, D., and Vesic, S. (2022). Evaluation of argument strength in attack graphs: foundations and semantics. *Artif. Intell.*, 302, 103607. doi: 10.1016/j.artint.2021.103607
- Amgoud, L., and Vesic, S. (2014). Rich preference-based argumentation frameworks. *Int. J. Approximate Reason.* 55, 585–606. doi: 10.1016/j.ijar.2013.10.010
- Andreas, J., Rosenthal, S., and McKeown, K. R. (2012). "Annotating agreement and disagreement in threaded discussion," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, eds N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odiijk, and S. Piperidis (Istanbul: European Language Resources Association), 818–822.
- Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., et al. (2017). Towards artificial argumentation. *AI Mag.* 38, 25–36. doi: 10.1609/aimag.v38i3.2704
- Baff, R. E., Wachsmuth, H., Khatib, K. A., and Stein, B. (2020). "Analyzing the persuasive effect of style in news editorial argumentation," in eds D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetrault *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020* (Association for Computational Linguistics), 3154–3160.
- Banerjee, P., and Barnwal, R. P. (2023). *Methods and Metrics for Explaining Artificial Intelligence Models: A Review*. Cham: Springer International Publishing.
- Baroni, P., Gabbay, D., Giacomin, M., and van der Torre, L. (2018). *Handbook of Formal Argumentation*. London: College Publications.
- Baroni, P., Rago, A., and Toni, F. (2019). From fine-grained properties to broad principles for gradual argumentation: a principled spectrum. *Int. J. Approximate Reason.* 105, 252–286. doi: 10.1016/j.ijar.2018.11.019
- Baroni, P., Romano, M., Toni, F., Aurisicchio, M., and Bertanza, G. (2015). Automatic evaluation of design alternatives with quantitative argumentation. *Argument Comput.* 6, 24–49. doi: 10.1080/19462166.2014.1001791
- Bench-Capon, T. (2003). Persuasion in practical argument using value-based argumentation frameworks. *J. Logic Comput.* 13, 429–448. doi: 10.1093/logcom/13.3.429
- Bench-Capon, T. (2012). "Open texture and argumentation: what makes an argument persuasive?" in *Logic Programs, Norms and Action*, eds A. Artikis, R. Craven, N. K. Çiçekli, B. Sadighi, K. Stathis (Berlin; Heidelberg: Springer-Verlag), 220–233.
- Berners-Lee, T., Hendler, J., and Lassila, O. (2001). *The Semantic Web*.
- Besnard, P., and Hunter, A. (2008). *Elements of Argumentation*. Cambridge, MA: The MIT Press.
- Bex, F., Lawrence, J., Snaith, M., and Reed, C. (2013). Implementing the argument web. *Commun. ACM* 56, 66–73. doi: 10.1145/2500891
- Black, E., and Hunter, A. (2008). "Using enthymemes in an inquiry dialogue system," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1 (AAMAS '08)* (Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems), 437–444.
- Boltuzic, F., and Snajder, J. (2016). "Fill the gap! analyzing implicit premises between claims from online debates," in *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016* (Berlin: The Association for Computer Linguistics).
- Bonzon, E., Delobelle, J., Konieczny, S., and Maudet, N. (2016). "A comparative study of ranking-Based semantics for abstract argumentation," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (Phoenix, AZ: AAAI Press), 914–920.
- Brewka, G., and Eiter, T. (2009). "Argumentation context systems: a framework for abstract group argumentation," in *Logic Programming and Nonmonotonic Reasoning, 10th International Conference, LPNMR 2009*, eds E. Erdem, F. Lin, and T. Schaub (Potsdam: Springer), 44–57.
- Budzynska, K., and Reed, C. (2011). "Speech acts of argumentation: inference anchors and peripheral cues in dialogue," in *Computational Models of Natural Argument, Papers from the 2011 AAAI Workshop* (San Francisco, CA).
- Cavoukian, A. (2013). *Privacy by Design: Leadership, Methods, and Results*. Dordrecht: Springer.

Acknowledgments

We confirm that we have retained the copyright for the material that originally appeared within Flouris et al. (2013, 2016).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Cerutti, F., Gaggi, S. A., Thimm, M., and Wallner, J. P. (2018). "Foundations of implementations for formal argumentation," in *Handbook of Formal Argumentation, also appears in IfCoLog Journal of Logics and their Applications*, Vol. 4, eds P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre (London: College Publications), 2623–2706.
- Chalaguine, L. A., and Hunter, A. (2019). "Knowledge acquisition and corpus for argumentation-based chatbots," in *Proceedings of the 3rd Workshop on Advances In Argumentation In Artificial Intelligence co-located with the 18th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2019)*, eds F. Santini and A. Toniolo (Rende), 1–14.
- Chalaguine, L. A., and Hunter, A. (2020). "A persuasive chatbot using a crowd-sourced argument graph and concerns," in *Computational Models of Argument - Proceedings of COMMA 2020*, eds H. Prakken, S. Bistarelli, F. Santini, and C. Taticchi (Perugia: IOS Press), 9–20.
- Charwat, G., Dvorák, W., Gaggi, S. A., Wallner, J. P., and Woltran, S. (2015). Methods for solving reasoning problems in abstract argumentation - A survey. *Artif. Intell.* 220, 28–63. doi: 10.1016/j.artint.2014.11.008
- Chen, S., Khashabi, D., Callison-Burch, C., and Roth, D. (2019). "PerspectroScope: a window to the world of diverse perspectives," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, eds M. R. Costa-jussà and E. Alfonseca (Florence: Association for Computational Linguistics), 129–134.
- Chesñevar, C., McGinnis, J., Modgil, S., Rahwan, I., Reed, C., Simari, G., et al. (2006). Towards an argument interchange format. *Knowl. Eng. Rev.* 21, 293–316. doi: 10.1017/S0269888906001044
- Cocarascu, O., and Toni, F. (2017). "Identifying attack and support argumentative relations using deep learning," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, eds M. Palmer, R. Hwa, and S. Riedel (Copenhagen: Association for Computational Linguistics), 1374–1379.
- Craandijk, D., and Bex, F. (2020). "Deep learning for abstract argumentation semantics," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, ed C. Bessiere (Yokohama), 1667–1673.
- da Costa Pereira, C., Tettamanzi, A., and Villata, S. (2011). "Changing one's mind: erase or rewind?" in *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, ed T. Walsh (Barcelona), 164–171.
- Dietz, E., Kakas, A., and Michael, L. (2022). Argumentation: a calculus for human-centric AI. *Front. Artif. Intell.* 5, 955579. doi: 10.3389/frai.2022.955579
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77, 321–357.
- Durmus, E., and Cardie, C. (2018). "Exploring the role of priodeep modular RNN approach for ethos mining beliefs for argument persuasion," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, eds M. A. Walker, H. Ji, and A. Stent (New Orleans, LA), 1035–1045.
- Duthie, R., and Budzynska, K. (2018). "A deep modular RNN approach for ethos mining," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, ed J. Lang (Stockholm), 4041–4047.
- Florou, E., Konstantopoulos, S., Koukourikos, A., and Karampiperis, P. (2013). "Argument extraction for supporting public policy formulation," in *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (Avignon: ACL)*, 49–54.
- Flouris, G., Bikakis, A., Patkos, T., and Plexousakis, D. (2013). *Globally Interconnecting Persuasive Arguments: The Vision of the Persuasive Web*. Technical report.
- Flouris, G., Bikakis, A., Patkos, T., and Plexousakis, D. (2016). "Argument extraction challenges in a new web paradigm," in *Dagstuhl Seminar 16161: Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments (Dagstuhl)*.
- Gabbay, D., Giacomin, M., Simari, G., and Thimm, M. (2021). *Handbook of Formal Argumentation*, Vol. 2. London: College Publications.
- Ghosh, D., Muresan, S., Wacholder, N., Aakhus, M., and Mitsui, M. (2014). "Analyzing argumentative discourse units in online interactions," in *Proceedings of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the Association for Computational Linguistics* (Baltimore, MD: The Association for Computer Linguistics), 39–48.
- Gordon, T. F., Prakken, H., and Walton, D. (2007). The Carneades model of argument and burden of proof. *Artif. Intell.* 171, 875–896. doi: 10.1016/j.artint.2007.04.010
- Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., et al. (2020). "A large-scale dataset for argument quality ranking: construction and analysis," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020* (New York, NY: AAAI Press), 7805–7813.
- Habernal, I., and Gurevych, I. (2016). "Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016* (Berlin: The Association for Computer Linguistics).
- Habernal, I., and Gurevych, I. (2017). Argumentation mining in user-generated web discourse. *Comput. Linguist.* 43, 125–179. doi: 10.1162/COLI_a_00276
- Hahn, U., and Oaksford, M. (2012). "Rational argument," in *The Oxford Handbook of Thinking and Reasoning*, eds K. Holyoak and R. Morrison (New York, NY: Oxford University Press), 277–298.
- Hidey, C., Musi, E., Hwang, A., Muresan, S., and McKeown, K. (2017). "Analyzing the semantic types of claims and premises in an online persuasive forum," in *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017* (Copenhagen: Association for Computational Linguistics), 11–21.
- Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., et al. (2021). "Knowledge graphs," in *Synthesis Lectures on Data, Semantics, and Knowledge* (Boston, MA: Springer), 1–237. doi: 10.2200/S01125ED1V01Y202109DSK022
- Hunter, A. (2013). A probabilistic approach to modelling uncertain logical arguments. *Int. J. Approximate Reason.* 54, 47–81. doi: 10.1016/j.ijar.2012.08.003
- Hunter, A. (2015). "Modelling the persuadee in asymmetric argumentation dialogues for persuasion," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015* (Buenos Aires: AAAI Press), 3055–3061.
- Khatib, K. A., Trautner, L., Wachsmuth, H., Hou, Y., and Stein, B. (2021). "Employing argumentation knowledge graphs for neural argument generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021* (Bangkok), 4744–4754.
- Kirschner, C., Eckle-Kohler, J., and Gurevych, I. (2015). "Linking the thoughts: analysis of argumentation structures in scientific publications," in *Proceedings of the 2nd Workshop on Argumentation Mining* (Denver, CO: Association for Computational Linguistics), 1–11.
- Kobbe, J., Opitz, J., Becker, M., Hulpus, I., Stuckenschmidt, H., and Frank, A. (2019). "Exploiting background knowledge for argumentative relation classification," in *2nd Conference on Language, Data and Knowledge, LDK 2019* (Leipzig), 8:1–8:14.
- Kuhlmann, I., and Thimm, M. (2019). "Using graph convolutional networks for approximate reasoning with abstract argumentation frameworks: a feasibility study," in *Scalable Uncertainty Management - 13th International Conference, SUM 2019* (Cogné: Springer), 24–37.
- Lagniez, J.-M., Lonca, E., Mailly, J.-G., and Rossit, J. (2021). Design and results of ICCMA 2021. *arXiv preprint arXiv:2109.08884*. Available online at: <https://arxiv.org/abs/2109.08884v1>
- Lawrence, J., and Reed, C. (2016). "Argument mining using argumentation scheme structures," in *Computational Models of Argument - Proceedings of COMMA 2016* (Potsdam: IOS Press), 379–390.
- Lawrence, J., and Reed, C. (2017). "Mining argumentative structure from natural language text using automatically generated premise-conclusion topic models," in *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017* (Copenhagen: Association for Computational Linguistics), 39–48.
- Lawrence, J., and Reed, C. (2019). Argument mining: a survey. *Comput. Linguist.* 45, 765–818. doi: 10.1162/coli_a_00364
- Lawrence, J., Snaith, M., Bex, F., and Reed, C. (2012). ArguBlogging, Arvina, and TOAST. *Front. Artif. Intell. Appl.* 245, 511–516. doi: 10.3233/978-1-61499-111-3-515
- Leite, J. A., and Martins, J. A. (2011). "Social abstract argumentation," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - IJCAI'11* (Barcelona: AAAI Press), 2287–2292.
- Lukin, S. M., Anand, P., Walker, M. A., and Whittaker, S. (2017). "Argument strength is in the eye of the beholder: audience effects in persuasion," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017* (Valencia: Association for Computational Linguistics), 742–753.
- Macagno, F. (2021). Argumentation schemes in AI: a literature review. Introduction to the special issue. *Argument Comput.* 12, 287–302. doi: 10.3233/AAC-210020
- Malmqvist, L., Yuan, T., Nightingale, P., and Manandhar, S. (2020). "Determining the acceptability of abstract arguments with graph convolutional networks," in *Proceedings of the Third International Workshop on Systems and Algorithms for Formal Argumentation co-located with the 8th International Conference on Computational Models of Argument (COMMA 2020)* (Perugia), 47–56.
- McGuinness, D., Lebo, T., and Sahoo, S. (2013). *PROV-o: The PROV Ontology*. W3C recommendation, W3C.
- Mercier, H., and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* 34, 57–111. doi: 10.1017/S0140525X10000968
- Misra, A., Anand, P., Tree, J. E. F., and Walker, M. A. (2015). "Using summarization to discover argument facets in online ideological dialog," in *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver, CO: The Association for Computational Linguistics), 430–440.

- Misra, A., Tandon, S., Ts, S., Anand, P., and Walker, M. (2017). "Summarizing dialogic arguments from social media," in *SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue* (Saarbrücken: ISCA).
- Nguyen, H. V., and Litman, D. J. (2018). "Argument mining for improving the automated scoring of persuasive essays," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)* (New Orleans, LO: AAAI Press), 5892–5899.
- Niculae, V., Park, J., and Cardie, C. (2017). "Argument mining with structured SVMs and RNNs," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, (Vancouver, BC: Association for Computational Linguistics), 985–995.
- Oraby, S., Reed, L., Compton, R., Riloff, E., Walker, M. A., and Whittaker, S. (2015). "And that's a fact: distinguishing factual and emotional argumentation in online dialogue," in *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015* (Denver, CO: The Association for Computational Linguistics), 116–126.
- Özsu, M. T. (2016). A survey of RDF data management systems. *Front. Comput. Sci.* 10, 418–432. doi: 10.1007/s11704-016-5554-y
- Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimaraes, G. A. S., Cruz, G. O. R., et al. (2023). "Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big Data Cogn. Comput.* 7, 1–31. doi: 10.3390/bdcc7010015
- Panisson, A. R., McBurney, P., and Bordini, R. H. (2021). A computational model of argumentation schemes for multi-agent systems. *Argument Comput.* 12, 357–395. doi: 10.3233/AAC-210555
- Patkos, T., Flouris, G., and Bikakis, A. (2016). "Symmetric multi-aspect evaluation of comments - extended abstract," in *ECAI 2016 - 22nd European Conference on Artificial Intelligence* (The Hague), 1672–1673.
- Peldszus, A., and Stede, M. (2013). From argument diagrams to argumentation mining in texts: a survey. *Int. J. Cogn. Inform. Nat. Intell.* 7, 1–31. doi: 10.4018/jcini.2013010101
- Peldszus, A., and Stede, M. (2015). "Joint prediction in MST-style discourse parsing for argumentation mining," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015* (Lisbon: The Association for Computational Linguistics), 938–948.
- Persing, I., and Ng, V. (2017). "Why can't you convince me? modeling weaknesses in unpersuasive arguments," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017* (Melbourne, VIC), 4082–4088.
- Prakken, H. (2009). "Models of persuasion dialogue," in *Argumentation in Artificial Intelligence* eds G. R. Simari and I. Rahwan (Boston, MA: Springer), 281–300.
- Prakken, H., Wyner, A. Z., Bench-Capon, T. J. M., and Atkinson, K. (2015). A formalization of argumentation schemes for legal case-based reasoning in ASPIC+. *J. Logic Comput.* 25, 1141–1166. doi: 10.1093/logcom/ext010
- Rach, N., Matsuda, Y., Daxenberger, J., Ultes, S., Yasumoto, K., and Minker, W. "Evaluation of argument search approaches in the context of argumentative dialogue systems," in *Proceedings of The 12th Language Resources and Evaluation Conference, LREC (2020)* (Marseille).
- Rahwan, I., Zablith, F., and Reed, C. (2007). Laying the foundations for a world wide argument web. *Artif. Intell.* 171, 897–921. doi: 10.1016/j.artint.2007.04.015
- Ranade, S., Gupta, J., Varma, V., and Mamidi, R. (2013). "Online debate summarization using topic directed sentiment analysis," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM 2013* (Chicago, IL: ACM), 7:1–7:6.
- Reed, C., Budzynska, K., Duthie, R., Janier, M., Konat, B., Lawrence, J., et al. (2017). The argument web: an online ecosystem of tools, systems and services for argumentation. *Philos. Technol.* 30, 131–160. doi: 10.1007/s13347-017-0260-8
- Reed, C., and Rowe, G. (2004). Araucaria: software for argument analysis, diagramming and representation. *Int. J. Artif. Intell. Tools* 13, 983. doi: 10.1142/S0218213004001922
- Reed, C., and Walton, D. (2005). Towards a formal and implemented model of argumentation schemes in agent communication. *Auton. Agents Multi Agent Syst.* 11, 173–188. doi: 10.1007/s10458-005-1729-x
- Reed, C., Wells, S., Devereux, J., and Rowe, G. (2008). "AIF+: dialogue in the argument interchange format," in *Computational Models of Argument: Proceedings of COMMA 2008* (Toulouse: IOS Press), 311–323.
- Rosenthal, S., and McKeown, K. (2015). "I couldn't agree more: the role of conversational structure in agreement and disagreement detection in online discussions," in *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Prague: The Association for Computer Linguistics), 168–177.
- Saint-Dizier, P. (2018). A two-level approach to generate synthetic argumentation reports. *Argument Comput.* 9, 137–154. doi: 10.3233/AAC-180035
- Schneider, J., Groza, T., and Passant, A. (2013). A review of argumentation for the social semantic web. *Seman. Web* 4, 159–218. doi: 10.3233/SW-2012-0073
- Schoop, M., de Moor, A., and Dietz, J. L. (2006). The pragmatic web: a manifesto. *Commun. ACM* 49, 75–76. doi: 10.1145/1125944.1125979
- Shneiderman, B. (1996). "The eyes have it: a task by data type taxonomy for information visualizations," in *Proceedings of the 1996 IEEE Symposium on Visual Languages* (Boulder, CO: IEEE Computer Society), 336–343.
- Skeppstedt, M., Peldszus, A., and Stede, M. (2018). "More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing," in *Proceedings of the 5th Workshop on Argument Mining, ArgMining@EMNLP 2018* (Brussels: Association for Computational Linguistics), 155–163.
- Slonim, N., Bilu, Y., Alzate, C., Bar-Haim, R., Bogin, B., Bonin, F., et al. (2021). An autonomous debating system. *Nature* 591, 379–384. doi: 10.1038/s41586-021-03215-w
- Spaans, J. P. (2021). "Intrinsic argument strength in structured argumentation: a principled approach," in *Logic and Argumentation - 4th International Conference, CLAR 2021* (Hangzhou: Springer), 377–396.
- Stab, C., Daxenberger, J., Stahlhut, C., Miller, T., Schiller, B., Tauchmann, C., et al. (2018). "ArgumenText: searching for arguments in heterogeneous sources," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 21–25 (New Orleans, LO: Association for Computational Linguistics).
- Stab, C., and Gurevych, I. (2014). "Annotating argument components and relations in persuasive essays," in *Proc. 25th Int. Conf. Computational Linguistics (COLING'14)* (Dublin), 1501–1510.
- Stab, C., and Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. *Comput. Linguist.* 43, 619–659. doi: 10.1162/COLL_a_00295
- Stede, M., and Schneider, J. (2018). *Argumentation Mining*. San Rafael, CA: Morgan & Claypool Publishers.
- Steenbergen, M., Bächtiger, A., Spörndli, M., and Steiner, J. (2003). Measuring political deliberation: a discourse quality index. *Comp. Eur. Polit.* 1, 21–48. doi: 10.1057/palgrave.ccp.6110002
- Sutterer, M., Droegehorn, O., and David, K. (2008). "UPOS: user profile ontology with situation-dependent preferences support," in *First International Conference on Advances in Computer-Human Interaction, ACHI 2008* (Sainte Luce), 230–235.
- Teruel, M., Cardellino, C., Cardellino, F., Alonso Alemany, L., and Villata, S. (2018). "Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki: European Language Resources Association).
- Thorn Jakobsen, T. S., Barrett, M., Sogaard, A., and Lassen, D. (2022). "The sensitivity of annotator bias to task definitions in argument mining," in *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022* (Marseille: European Language Resources Association), 44–61.
- Trautmann, D., Fromm, M., Tresp, V., Seidl, T., and Schütze, H. (2020). Relational and fine-grained argument mining. *Datenbank-Spektrum* 20, 99–105. doi: 10.1007/s13222-020-00341-z
- van Gelder, T. (2007). The rationale for Rationale. *Law Probabil. Risk* 6, 23–42. doi: 10.1093/lpr/mgm032
- Vassiliades, A., Bassiliades, N., and Patkos, T. (2021). Argumentation and explainable artificial intelligence: a survey. *Knowledge Eng. Rev.* 36, e5. doi: 10.1017/S0269888921000011
- Vassiliades, A., Patkos, T., Bikakis, A., Flouris, G., Bassiliades, N., and Plexousakis, D. (2020). "Preliminary notions of arguments from commonsense knowledge," in *SETN 2020: 11th Hellenic Conference on Artificial Intelligence* (Athens: ACM), 211–214.
- Verheij, B. (2003). Dialectical argumentation with argumentation schemes: an approach to legal logic. *Artif. Intell. Law* 11, 167–195. doi: 10.1023/B:ARTI.0000046008.49443.36
- Villata, S., Boella, G., Gabbay, D. M., and van der Torre, L. W. N. (2013). A socio-cognitive model of trust using argumentation theory. *Int. J. Approximate Reason.* 54, 541–559. doi: 10.1016/j.ijar.2012.09.001
- von Ahn, L., and Dabbish, L. (2008). Designing games with a purpose. *Commun. ACM* 51, 58–67. doi: 10.1145/1378704.1378719
- Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., et al. (2017a). "Computational argumentation quality assessment in natural language," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017* (Valencia: Association for Computational Linguistics), 176–187.
- Wachsmuth, H., Potthast, M., Al-Khatib, K., Ajjour, Y., Puschmann, J., Qu, J., et al. (2017b). "Building an argument search engine for the web," in *Proceedings of the 4th Workshop on Argument Mining* (Copenhagen: Association for Computational Linguistics), 49–59.
- Walton, D. (2006). *Fundamentals of Critical Argumentation*. Cambridge, UK: Cambridge University Press.
- Wyner, A. Z. (2016). A functional perspective on argumentation schemes. *Argument Comput.* 7, 113–133. doi: 10.3233/AAC-160010
- Zografistou, D., Flouris, G., Patkos, T., and Plexousakis, D. (2018). "Implementing the ArgQL query language," in *Computational Models of Argument - Proceedings of COMMA 2018* (Warsaw: IOS Press), 241–248.



OPEN ACCESS

EDITED BY

Loizos Michael,
Open University of Cyprus, Cyprus

REVIEWED BY

John Zeleznikow,
La Trobe University, Australia
Markus Ulbricht,
Leipzig University, Germany

*CORRESPONDENCE

Antonino Rotolo
✉ antonino.rotolo@unibo.it
Giovanni Sartor
✉ giovanni.sartor@unibo.it

RECEIVED 23 December 2022

ACCEPTED 11 August 2023

PUBLISHED 04 September 2023

CITATION

Rotolo A and Sartor G (2023) Argumentation
and explanation in the law.
Front. Artif. Intell. 6:1130559.
doi: 10.3389/frai.2023.1130559

COPYRIGHT

© 2023 Rotolo and Sartor. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Argumentation and explanation in the law

Antonino Rotolo^{1*} and Giovanni Sartor^{1,2*}

¹Alma AI and Department of Legal Studies, Alma Mater Studiorum - Università di Bologna, Bologna, Italy,

²Department of Law, European University Institute, Florence, Italy

This article investigates the conceptual connection between argumentation and explanation in the law and provides a formal account of it. To do so, the methods used are conceptual analysis from legal theory and formal argumentation from AI. The contribution and results are twofold. On the one hand, we offer a critical reconstruction of the concept of legal argument, justification, and explanation of decision-making as it has been elaborated in legal theory and, above all, in AI and law. On the other hand, we propose some definitions of explanation in the context of formal legal argumentation, showing a connection between formal justification and explanation. We also investigate the notion of stable normative explanation developed elsewhere in Defeasible Logic and extend some complexity results. Our contribution is thus mainly conceptual, and it is meant to show how notions of explanation from literature on explainable AI and legal theory can be modeled in an argumentation framework with structured arguments.

KEYWORDS

explanation, argumentation, legal reasoning, defeasibility, normative systems, justification

1. Introduction

Argumentation is critically relevant to law, whose application involves deliberation over the ascertainment of uncertain past facts, as well as the interpretation and application of general legal rules to particular cases, in consideration of relevant values and principles.¹

Legal problem solving involves dialectical and indeed adversarial interactions in which different ways of reasoning are deployed: probabilistic, deductive and presumptive inferences, the use of analogies, appeals to precedent and policy, and the balancing of interests.

Legal decisions have the authority to be coercively enforced, as issuing from the political community. Thus, such decisions need to be justified: reasons must be provided of why certain claims were endorsed, based on what reasons, and it must be specified why such reasons prevailed over the reasons to the contrary. These justifications need to be critically evaluated, to determine whether they succeed in explaining legal decision in a way that is satisfactory for the individuals involved and for the society.

While legal theory has extensively studied legal argumentation (see [Perelman and Olbrechts-Tyteca, 1969](#); [MacCormick, 1978](#); [Alexy, 1989](#)), a formal account of it has only been provided by the AI & Law research, which has profited from, and contributed to, the logical tools for argumentation made available within AI (for an overview, see [Prakken and Sartor, 2015](#)). In fact, AI & Law researchers have not only applied AI-based argumentation techniques to the law, but have also made innovative contributions to the development of formal models of argumentation.

An open research question, which is drawing more and more attention in the literature, is the conceptual and formal investigation of the relation between justification and explanation of legal decisions, especially, when norms are crucial in the reasoning process. This article will mainly address this issue.

¹ The introduction and parts of Section 3 elaborate on materials from [Prakken and Sartor \(2015\)](#).

1.1. Purpose and synopsis of this contribution

We believe there is a still overlooked research challenge, which, taking stock of major achievements in legal theory, concerns the relationship between the “justification of legal arguments” and the “explanation of normative conclusions.” To tackle this issue, we aim to connect two research domains, employing formal argumentation: the investigation of AI & Law, which focuses on justifying (automated) legal decision-making, and the examination of explanations within the context of eXplainable AI.

Our contribution is primarily conceptual, aiming to demonstrate how ideas proposed in works such as the one by Miller (2019) or explored in legal theory can be represented within an argumentation framework.

In the light of the importance of argumentation for the legal domain, this article thus aims at contributing to the following related aspects:

- Reconstructing, from the AI & Law literature, the main models of legal argument and formal argumentation, and linking these models to the concepts of justification and explanation;
- Given the above conceptual background, proposing some definitions of explanation in the context of formal legal argumentation.

The layout of the article is as follows. Section 2 clarifies the distinction in the law between justification and explanation. Section 3 develops as follows: after recalling why the law is an argumentation framework (Section 3.1), we will discuss the need to provide explanations when norms are used as preconditions for inferring and issuing other norms (Section 3.2) or for applying them (Section 3.3). We will then consider applications in legal interpretation (Section 3.4) and in case-based reasoning (Section 3.5). Sections 4, 5 offer a conceptual analysis of legal explanation in formal argumentation: the building blocks are recalled in Section 4, while Section 5 presents some definitions of the idea explanation in legal argumentation and investigates the concept of stable explanation extending previous work. Section 6 clarifies the originality of our contributions and discusses related and future work. Some conclusions end the paper.

2. Justification, explanation, and argumentation in legal reasoning

In this section we shall discuss how argumentation has a foundational role with regard to the justification and explanation of normative conclusions.

2.1. Justification and explanation in legal decision-making

An extensive discussion of the relation between normative explanation and justification (Baier, 1958, chap. 6) is beyond the scope of this paper. Let us just remark that, while a vast

literature exists on the concept of an explanation in philosophy (Achinstein, 1983; Pitt, 1988) legal theory has mainly focused on justification, taking this concept as central in the context of legal decision-making (Alexy, 1989; Peczenik, 1989). From the legal theory perspective, it may seem that explanations are a byproduct of justifications: the arguments justifying a decision, on the basis of facts and norms, also provide an explanation of the same decision.

The connection between explanation and justification has also emerged in AI, where more attention has been devoted to the concept of explanation, especially in the debate on eXplainable AI (XAI) (Miller et al., 2022). The AI & Law community has also worked toward explanation, since both “transparency” and “justification” of (automated) legal decision-making require providing explanations (Atkinson et al., 2020; Governatori et al., 2022c; Prakken and Ratsma, 2022).

Legal decision-making (and consequently, also legal advice) is a complex multi-step process that involves addressing factual and normative issues, based on empirical evidence and legal questions. Different answers to such issues are often possible, depending on the ethical and political preferences and the psychological attitudes of the decision-makers. The extent to which such preferences and attitudes may determine the outcome of the case is constrained by the available evidence and applicable norms. However, a space for discretion, broadly understood, remains, and this space is adjustable, since constraints themselves are to be interpreted by the decision-makers, according to their view of the role of decision-maker (typically judges) and of the principle of the separation of powers.

Within an argumentation-based approach, *the justification of a legal decision may be viewed as an argument structure aimed to show that the decision is right or correct, according to a convincing reconstruction of facts and norms*. Justifications are pervasive in the law, since, as noted above, legal decision-makers are usually required to publicly provide rational grounds for the normative correctness of their decisions (at least for important ones). Justifications may also be produced, possibly integrating the original ones, at a subsequent time, by those who agree with such decisions and want to provide further reasons supporting them.

Consider for instance *Dobbs v. Jackson Women’s Health Organization*, 19-1392 U.S. 597 (2022) decision by the US Supreme Court, which denied the existence of a constitutional right to abortion, contrary to the previous *Roe v. Wade*, 410 U.S. 113 (1973) decision, which had affirmed that right. The majority of the *Dobbs* judges provided a justification of that decision based on certain legal doctrines on the interpretation of the US constitution (a version of the so-called originalism), on federalism, on the separation of powers, which require according to their view that the legality of abortion is decided at the State level, rather than at the federal level. More extended justifications of that decisions have been provided by legal scholars who agree with its content and want to support its correctness with further considerations. On the other hand, the judges in the dissenting opinion strongly criticized this justification, and so did scholars and activists opposing the *Dobbs* decision.

In legal theory it is common to distinguish the “discovery” process through which decision-makers endorse certain conclusions on the relevant issues—guided by the information they access, but also by their intuitions and by their tacit expert

knowledge—and the process of building an accessible justification of that decision, which may appear convincing or at least acceptable to the parties and the public (MacCormick, 1978). Justification usually follows discovery, and selectively uses the information elicited during discovery, in order to provide a rhetorically effective account. However, dialectical interactions between the two processes exist: on the one hand considerations developed during the discovery process may enter into the justification, on the other hand the necessity to build a convincing justification may guide the process of discovery, leading the decision-makers to reject or amend the outcomes for which a convincing justification could not be found.

It seems to us that in any case a description of the discovery process is no substitute for a justification as just described: first of all, many aspects of the process of discovery are not accessible to description, pertaining to the unconscious working of the decision-maker's mind; secondly, some moves in the discovery process may pertain to taking wrong directions, or anyway to aspects that are not relevant for the goal of providing a publicly acceptable justification. On the other hand, however, certain inference steps that took place during discovery (including logical and statistical inferences, the assessment of competing factors, the interaction of rules and exceptions, presumption, etc.) can be recovered for the purpose of building a justification.

2.2. Types of legal explanation: conceptual distinctions

While *justifications are reasoned defenses of (legal) decisions* by the authors of such decisions or third parties supporting the same decisions, *explanations involve a third-party perspective*, which does not presuppose the endorsement of the explained decisions (for a general philosophical discussion, see Davidson, 1963).

We may indeed distinguish two ways of explaining legal decisions: *causal explanations*, and *rational reconstructions*.

Causal explanations of legal decisions aim at identifying social, ideological, or political factors that contribute to the outcomes of legal cases, inducing decision-makers to adopt such outcomes. For instance, in the Dobb case we might consider that the outcome was determined by the political position of the majority of the judges (positioned in the right-wing side and nominated by republican presidents), their religious convictions, their ideological commitments, their connections with certain groups of the population, etc. In some cases, the causal explanation may include pointing to failures in the decisional process: the decision-makers were affected by their prejudices, were bribed, their decision was instrumental to favoring their friends or harm their enemies, etc.

This *extra-legal and extra-systemic explanation* of legal decisions can be distinguished from the *intra-legal and intra-systemic rational explanation* (i.e., a rational reconstruction), by which we may understand the attempt to identify reasons why certain decisions may be legally appropriate, given the beliefs, view-point and political-ethical-legal commitments of those who support such decisions (and first of all of the decision-makers who adopted them). A broad notion, which fits with our analysis, is proposed by Väyrynen (2021) for whom normative explanations

are “explanations of why things are wrong, good, or unfair.” In the context of legal decision-making, we may say that a normative explanation is an account of why a legal evaluation (on the legality, illegality of action, the ascription of rights or obligations) is considered to be correct on the basis of *both norms and facts* (a combination that was first emphasized by Schroeder, 2005).

Rational explanations, as well as justifications can take the shape of an argumentation framework, in which, besides presenting the arguments favoring the decision, arguments to the contrary are considered and defeated. This perspective involves a “principle of charity,” in the sense that it is assumed that the decision is the outcome of reasoned factual and legal considerations, even though we may disagree with the substance of such considerations. Thus, those who disagree with the Dobbs decision, can still provide a rational explanation (reconstruction) of that decision by presenting a coherent narrative including legally relevant reasons in favor of that decision, together with the assessment of such reasons according to the perspective of those who endorse them. Nevertheless, the opponents of the same decision may continue to consider that it was wrong, since stronger reasons, according to their perspective, exist for reaching the opposite conclusion. The opponent of Dobbs can also merge their critical considerations with the rational explanation of the decision they disagree with. In such a case, a critical argumentation framework is obtained, in which the arguments explaining the decision are defeated by the arguments against that decision (consider for instance, a legal essay that attacks the conclusion as well as the reasoning of the judges in the Dobbs case).

2.3. Our conceptual standpoint

In conclusion, it seems to us that both the justification of a legal decision and its rational explanation, as described above, can be captured through argumentation. Both ideas presuppose that an outcome (the decision of a case) should be supported by reasons, and that these reasons should prevail over the reason to the contrary, according to a certain perspective. The *distinction between justification and rational explanation, according to our analysis, pertains to pragmatics, rather than to semantics*. It concerns the purpose of the exercise: providing support to a decision we endorse (being those who propose it, or aim to advance or defend it) or rather accounting for the support that is attributed to a decision by those who endorse it, for reasons endorsed by them. In the following, when speaking of explanations, without further clarifications, we cover both justifications and rational explanations.

The distinction between justification and (rational) explanation thus seems to rely on a perspectival approach. For an agent a_1 : (a) a decision d (by a_1 or by other agents) is justified iff it is supported by prevailing reasons in the context of the attitudes and beliefs of a_1 ; (b) a decision d by an agent a_2 is (rationally) explained if it is supported by prevailing reasons in the context of the attitudes and beliefs of a_2 .

In the context of the application of AI technologies to legal decision-making the relation between discovery (the activity through which a system constructs an answer to a legal issue) and

justification-explanation (the attempt to provide reasons for that answer) tends to take a different form in knowledge-based systems (including formal-argumentation system), on the one hand, and in opaque machine learning applications, on the other hand. In the first case, the argumentative justification-explanation of a case can be constructed on the basis of the very inferences and reasoning patterns through which the system came to determine its outcome; in the second case an argumentative justification-explanation has to be constructed as a parallel exercise, meant to mimic the opaque inference of the system. In both cases, the construction of argumentative explanations presupposes the availability of a knowledge base of rules and concepts, from which arguments can be constructed.

In this paper, we shall assume that such a knowledge-base is available, and we shall consider to what extent it can be used to build argumentation frameworks. Given an argumentation framework we shall consider, by deploying an argumentation semantics, what arguments and conclusions are supported by that framework, where this notion of support may be viewed as a kind of justification: an outcome is justified by the (grounded) extension or labeling in which it is included. Based on this idea, we shall provide some notions that clarify aspects of legally relevant explanations.

First, we shall discuss whether an explanation can be viewed as an argument set that is suitable to support the explanandum (within the given argumentation framework): if any arguments in the set were not available the explanandum would not be derived, through that explanation.

Focusing especially on factual premises and norms, we shall then consider contrastive explanations, which elicit, under minimality conditions, those facts or norms whose presence or removal would preempt the derivation of the explanandum.

We are aware that our analysis cannot cover all aspects that are addressed under philosophical conceptions of a (normative) explanation, but we believe we will provide a sufficiently rich account that makes an essential use of the distinctive elements the legal knowledge base.

3. Models of legal argument

The adoption of argumentative model for the justification-explanation of legal decisions was motivated by the fact that purely deductive approaches fail to capture key aspects of legal reasoning, such as conflicts between competing rules, the relation between rules and exceptions, the significance of factors, interpretive and case-based reasoning, and more generally, the dialectical and adversarial nature of legal interactions (Perelman and Olbrechts-Tyteca, 1969; Alexy, 1989; Walton et al., 2008; Bongiovanni et al., 2018).

Argumentation pervades all the three dimensions of the law distinguished by Hart (1994):

- **Norm recognition and hierarchies:** legal systems consist of norms and provide criteria for establishing whether any norms belong to them; legal systems assign to their norms a different ranking status and organize them in hierarchies (e.g., constitutional norms are stronger than legislative acts);
 - **Norm change:** legal systems change and include criteria governing their dynamical evolution;
 - **Norm application:** the norms in a legal system are applied to concrete cases, and this process is based on interpretive and procedural criteria specified by that system.
- In a reasoning and argumentative perspective, we can think of the above dimensions as follows:*
- Arguments can be used for inferring, issuing, or adopting norms, and for determining how norms are related with one another (e.g., for establishing when one norm may override another one in case of conflicts; **Norm recognition and hierarchies**);
 - Arguments can be used for proposing and implementing revisions to legal systems (**Norm change**);
 - Arguments can be used for advancing interpretations of legal provisions, supporting them against alternative interpretations (e.g., when different interpretive canons, as applied to the same provision, offer different legal solutions for the same case) and for applying the resulting norms (**Norm application**).
- In the following, we briefly recall the main contribution of argumentation theory in AI & Law to some of these dimensions and identify some challenges to be addressed in regard to the distinction between justification and explanation.

3.1. The law as an argumentation framework

It has been argued that the law itself can be described as a complex argumentation framework (Prakken and Sartor, 2015). Under this general assumption, arguments must determine (and thus explain) the way in which *norms interplay with one another in legal systems* (Alchourron and Bulygin, 1971). Defeasible argumentation (Dung, 1995; Pollock, 1995) has indeed been used to address conflicts between norms and ways to resolve such conflicts through meta-arguments, as well as the interactions between legal rules and the reasons supporting them (Hage, 1997; Prakken and Sartor, 2015).

Through formal argumentation, among others, the following challenges can be addressed:

- Explaining the interplay of legal norms. When there is a conflict of legal rules r_1 and r_2 , both applicable to the case at hand, then a decision for r_1 's outcome must include a preference for r_1 and possibly reasons for that preference.
- Explaining the application of norms. In deciding a case a procedure has to be followed where facts have to be assessed in compliance with legal constraints, rules have to be identified and their applicability assessed.
- Explaining the interpretation of legal norms. When alternative interpretations i_1 and i_2 of a legal provision exist, then a decision for the outcome corresponding to i_1 must be

supported by the reasons why i_1 rather than i_2 should be accepted as the interpretation of that provision.

3.2. Explaining the interplay of legal norms

Let us first consider the need to provide explanations where norms are used for inferring, issuing, or applying other norms.

Assume that norms in the legal system L are represented as rules of the form $r: \phi_1, \dots, \phi_n \Rightarrow \psi$ (where r is the name of the norm). Then a preference relation $>$ can capture a hierarchy over L that enables collisions between norms being addressed. Consider for example

$$L = \{ \{ r: \phi_1, \dots, \phi_n \Rightarrow \psi, \quad s: \psi \Rightarrow \pi, \quad t: \omega \Rightarrow \neg\pi \} \\ > = \{ \langle s, t \rangle \} \}.$$

Assume also that the antecedents of r and t , i.e., facts $\phi_1, \dots, \phi_n, \omega$ are the case. Because s is hierarchically superior to t , then an argument A concatenating ϕ_1, \dots, ϕ_n, r and s successfully supports the conclusion π , defeating the argument concatenating ω and t . Jurists usually would say that this argument legally grounds and justifies π in L . Notice that the law “is not concerned with the absolute rationality of the normative statement in question, but only with showing that it can be rationally justified within the framework of the validly prevailing legal order” (Alexy, 1989, p. 220). This simple context illustrates different legally relevant explanations of π . We may say that conclusion π is explained:

- By argument A , which grounds conclusion π upon the relevant facts;
- By the whole of L plus all facts of the case, which together provide for the conflicting arguments and for the preference solving their conflict;
- By each fact in ϕ_1, \dots, ϕ_n , since one may counterfactually argue that without any of them we would have $\neg\pi$ rather than π ;
- By each of the rules r and s , since without either of them π could not be (sceptically) inferred;
- By the preference $s > t$, without which also π could not be inferred.

3.3. Explaining the application of the law

When the law is applied to cases (e.g., by judges in courts), legal theory traditionally breaks down the analysis of judicial decisions into three dimensions: the so-called question of fact (*quaestio facti*), i.e., reconstructing the facts of the case on the basis of the available evidence, the ways in which proceedings develop (judicial procedures), and the so-called question on law (*quaestio juris*), i.e., interpreting the law to identify the applicable legal rule. Within AI & Law, an in-depth analysis has been developed of evidential reasoning, comparing different approaches to it (Verheij et al., 2016). The procedural aspects of decisions have been investigated in regard to ideas such as the standard of proofs, presumptions, and burdens of proof (Prakken and Sartor, 2006; Calegari and

Sartor, 2021; Kampik et al., 2021). Formalizations have also been developed for protocols governing the admissibility and impact of arguments in legal debates (Gordon, 1995; Governatori et al., 2014). More recently, the idea that multiple argument schemes can be used in legal arguments has been explored, as well as the issue of which argumentative strategies are most effective in different legal disputes from a game-theoretical perspective (Roth et al., 2007; Riveret et al., 2008).

Jurists naturally resort to *causal explanation* in the context of reasoning about evidence (Walton, 2002), where competing accounts of the facts of the case are developed on the basis of the available evidence. In this domain, AI & Law research has devoted an extensive effort and discussed classic issues such as the relation between abductive and counterfactual reasoning and legal argumentation (see, again Prakken and Sartor, 2015 for an overview of the literature, see Liepina et al., 2020 for a recent attempt to identify causal argument schemes for causal reasoning).

Logical models have also been used to relate legal norms to the cases at hand and explain why such norms are applicable to the given facts. One framework that has been developed for this purpose is called reason-based logic (RBL), which focuses on how principles, goals, and rules can influence the interpretation of legal provisions (Hage, 1997).

3.4. Explaining the interpretation of the law

Legal interpretation has been viewed as a decision-making problem, in which the goal is to choose the best interpretation based on its consequences for promoting and demoting values (Atkinson and Bench-Capon, 2007; da Costa Pereira et al., 2017). Another approach is the argument-scheme approach, which considers interpretive canons using defeasible rules to interpret legal provisions and resolving conflicts by comparing the reasons behind different interpretations (Rotolo et al., 2015; Walton et al., 2021). The latter idea fits legal theories that view interpretive canons as reasoning patterns for constructing arguments aimed at justifying interpretive outcomes. Examples of canons by McCormick and Summers (1991) are:

Argument from ordinary meaning: if a statutory provision can be interpreted according to the meaning a native speaker of a given language would ascribe to it, it should be interpreted in this way, unless there is a reason for a different interpretation.

Argument by coherence: a provision should be interpreted in light of the whole statute it is part of, or in light of other provisions it is related to.

Teleological argument: a provision should be interpreted as applied to a particular case in a way compatible with the purpose that the provision is supposed to achieve.

Arguments from general principles: whenever general principles, including principles of law, are applicable to a provision, one should favor the interpretation that is most in conformity with these general legal principles.

According to Rotolo et al. (2015) and Walton et al. (2021), the structure of interpretive arguments can be analyzed using *interpretation rules*, where the antecedent of interpretation rules can be of any type, while the conclusion is an interpretive act I of a provision n leading to an interpretive result ψ for n which expresses such an interpretation paraphrasing n into ψ . An example of an interpretation rule is the following:

$$r' : \phi_1, \dots, \phi_n \Rightarrow I_{\text{teleological}}(n_1^L, \psi) \quad (1)$$

Rule r' states that, if ϕ_1, \dots, ϕ_n hold, then the interpretive canon to be applied in legal system L for provision n_1 is the teleological interpretation, which returns ψ .

Now suppose to have the following rules (the example logically mirrors the one in Section 3.2):

$$\begin{aligned} R = \{ & \{r' : \phi_1, \dots, \phi_n \Rightarrow I_{\text{teleological}}(n_1^L, \psi) \\ & s' : I_{\text{teleological}}(n_1^L, \psi) \Rightarrow I_{\text{coherence}}(n_2^L, \pi), \\ & t' : \Rightarrow I_{\text{ordinary}}(n_2^L, \neg\pi)\} \\ & \Rightarrow \{(s', t')\} \}. \end{aligned}$$

In legal theory, we may say that the interpretation of n_2 as π is justified in the legal system L (on modeling interpretation through argumentation, see Walton et al., 2021; Sartor, 2023). We may also say that the argument built with r' and s' explains this outcome, or, also, that ϕ_1, \dots, ϕ_n explain it.

3.5. Explaining the use of judicial cases

Legal systems often rely on past cases to guide decision-making and legal reasoning. A popular AI & Law approach to case-based reasoning consists in focusing on factors, namely, on features of cases that favor or disfavor certain outcomes (Rissland and Ashley, 1987; Ashley, 1990; Ashley and Aleven, 1991). The presence or absence of certain factors in a new case, or in precedents cases, can be used to support or challenge legal claims. There have been various developments of the factor-based approach within AI & Law, including the use of multivalued factors (Bench-Capon and Rissland, 2002) and hierarchies of factors (Aleven and Ashley, 1997), as well as logical mechanisms for determining when a decision is consistent or inconsistent with a case base (Horty, 2011).

Investigations have been developed on the combination of models of case-based reasoning with formal approaches to defeasible argumentation (Berman and Hafner, 1993; Bench-Capon and Sartor, 2003; Bench-Capon et al., 2013; Maranhão et al., 2021). Accordingly, a case can be reconstructed as expressing two competing rules and a preference for one of them (Prakken and Sartor, 1998): the conjunction of the factors ϕ_1, \dots, ϕ_n which are present in the case and support its outcome ψ corresponds to a defeasible rule $\phi_1, \dots, \phi_n \Rightarrow \psi$, which prevails over the rule $\chi_1, \dots, \chi_m \Rightarrow \neg\psi$, whose antecedent is the conjunction of all factors χ_1, \dots, χ_m in the case which support the outcome $\neg\psi$. The rules involved in factor-based reasoning are defeasible in that new factors can explain deviations from earlier decisions.

In Liu et al. (2022a) case-based reasoning and classifier systems are connected, and on this basis different kinds of case-based

explanations are defined such as abductive and contrastive ones. The logic of Liu et al. (2022a) is based on modal logic and does not directly capture the argumentative nature of case-based reasoning, as recalled above. Prakken and Ratsma (2022) uses argumentation—based on multi-valued factors (dimensions)—to explain the outcome of legal cases.

4. Formal argumentation

In this section we present formal argumentation and illustrate its application to legal reasoning. Argumentation frameworks have been proposed by Dung (1995) to investigate the general aspects of dialectical reasoning without specifying the internal structure of arguments. Many semantic models have been developed (Baroni and Giacomin, 2009) for abstract argumentation. Such models determine what arguments can be accepted, by considering not only how such arguments directly conflict with each other, but also how arguments can be indirectly defended by other arguments. Among them, several options have been acknowledged as appropriate in legal reasoning (see Prakken and Sartor, 2023). However, since we work in this paper on argumentation for reasoning with norms, we follow Governatori et al. (2021) and Governatori and Rotolo (2023). These works suggest that when norms collide and no priority principles can apply (such as the principles *lex superior*, *lex posterior* and *lex specialis*), a skeptical approach may be the most appropriate one, especially when legal effects of norms are obligations or sanctions. For the sake of simplicity, we focus on grounded semantics.

Let us first of all recall from the literature some basic formal concepts.

Definition 1 (Argumentation framework and semantics).

Argumentation framework. An argumentation framework AF is a pair (\mathcal{A}, \gg) where \mathcal{A} is a set of arguments, and $\gg \subseteq \mathcal{A} \times \mathcal{A}$ is a binary, attack relation.

Conflict-free set. A set S of arguments is said to be conflict-free if, and only if there are no arguments A and B in S such that B attacks A .

Argument defense. Let $S \subseteq \mathcal{A}$. The set S *defends* an argument $A \in \mathcal{A}$ if, and only if for each argument B attacking A there is an argument $C \in S$ that attacks B .

Complete extension. Let $AF = (\mathcal{A}, \gg)$ and $S \subseteq \mathcal{A}$. S is a complete extension of AF if and only if S is conflict-free and $S = \{A \in \mathcal{A} \mid S \text{ defends } A\}$.

Grounded extension. A grounded extension $GE(AF)$ of an argumentation framework AF is the minimal complete extension of AF.

Justified argument and conclusion. An argument A and its conclusion $\text{Conc}(A)$ are justified w.r.t. an argumentation framework AF if, and only if $A \in GE(AF)$.

Rejected argument and conclusion. An argument A and its conclusion $\text{Conc}(A)$ are rejected w.r.t. an argumentation framework AF is, and only if $A \notin GE(AF)$.

While abstract argumentation is not concerned with the internal structure of arguments, it was argued in the AI & Law literature the importance of devising argumentation frameworks where arguments have a logical structure (see Sartor, 2005; Prakken and Sartor, 2015; Governatori et al., 2021). If the underlying language of an argumentation framework refers to any logic L , arguments can roughly correspond to proofs in L (Prakken and Vreeswijk, 2002). As done by Governatori et al. (2004), Prakken (2010), and Toni (2013), given the above framework the (internal) logical structure of arguments can be specified using rule-based systems in such a way that rules correspond, e.g., to norms or normative reasoning patterns (such as in the case of interpretation rules) (Sartor, 2005; Prakken and Sartor, 2015; Governatori et al., 2021) and arguments are logical inference trees built from them.

Definition 2 (Language). The language consists of *literals* and *defeasible rules*. Given a set PROP of propositional atoms, the set of literals is $\text{Lit} = \text{PROP} \cup \{\neg p \mid p \in \text{PROP}\}$. We denote with $\sim\phi$ the *complementary* of literal ϕ ; if ϕ is a positive literal ψ , then $\sim\phi$ is $\neg\psi$, and if ϕ is a negative literal $\neg\psi$, then $\sim\phi$ is ψ .

Let Lab be a set of unique rule labels. A *defeasible rule* r with $r \in \text{Lab}$ has the form $\text{Ant}(r) \Rightarrow \text{Head}(r)$, where

- $\text{Ant}(r)$, called the *antecedent* or the *premises* of r , is a subset of Lit (which may be empty) and
- $\text{Head}(r)$ is a literal in Lit, called the *consequent* or *head* of r .

If R is a set of rules,

- $R[\phi]$ is the set of rules in R with head ϕ ,
- $\text{ANT}(R)$ is the union of all antecedents of all rules in R (i.e., it contains all literals in the antecedents of such rules).

Any defeasible rule whose antecedent is satisfied provides sufficient support to its conclusion unless there is evidence contrary to that conclusion.²

Following Governatori et al. (2004) we use the term argumentation theory to denote the rule-based knowledge from which argumentation frameworks are built. Notice that, as done by Antoniou et al. (2001), we distinguish a set of indisputable statements called *facts*, even though, without loss of generality, we impose some restrictions on it to keep things simpler.

Definition 3 (Argumentation theory). An *argumentation theory* D is a structure

$$(R, F, >)$$

where

- R is a (finite) set of defeasible rules,
- $F \subseteq \text{Lit}$ is a consistent set of indisputable statements called *facts* such that, for each $\phi \in F$, $R[\phi] \cup R[\sim\phi] = \emptyset$, and

² In several systems other two kinds of rules are allowed: strict rules and defeaters. A strict rule is a rule in the classical sense: whenever the antecedent holds, so indisputably is the conclusion. A defeater is a rule that cannot be used to draw any conclusion, but can provide contrary evidence to complementary conclusions.

- $> \subseteq R \times R$ is a binary relation on R called *superiority relation*.

The relation $>$ describes the relative strength of rules, that is to say, when a single rule may override the conclusion of another rule; it is required to be irreflexive, asymmetric and acyclic (i.e., its transitive closure is irreflexive).

By combining the rules in a theory, we can build arguments [we adjust the definition by Prakken (2010) to meet Definition 3]. Let us first introduce some notation: for a given argument A , $\text{Conc}(A)$ returns A 's conclusion, $\text{Sub}(A)$ returns all its sub-arguments, $\text{Rules}(A)$ returns all the rules in the argument and, finally, $\text{TopRule}(A)$ returns the last inference rule in A .

Definition 4 (Argument). Let $D = (R, F, >)$ be an argumentation theory. An argument A for ϕ constructed from D has either the form $\Rightarrow_F \phi$ (*factual argument*), where $\phi \in F$, or the form $A_1, \dots, A_n \Rightarrow_r \phi$ (*plain argument*), where $1 \leq k \leq n$, and

- A_k is an argument constructed from D , and
- $r : \text{Conc}(A_1), \dots, \text{Conc}(A_n) \Rightarrow \phi$ is a rule in R .

With regard to a factual argument $\Rightarrow_F \phi$:

$$\text{Conc}(A) = \phi; \quad \text{Sub}(A) = \emptyset; \quad \text{TopRule}(A) = \emptyset; \quad \text{Rules}(A) = \emptyset$$

With regard to a plain argument $A = A_1, \dots, A_n \Rightarrow_r \phi$:

$$\begin{aligned} \text{Conc}(A) &= \phi \\ \text{Sub}(A) &= \text{Sub}(A_1), \dots, \text{Sub}(A_n), A \\ \text{TopRule}(A) &= r : \text{Conc}(A_1), \dots, \text{Conc}(A_n) \Rightarrow_r \phi \\ \text{Rules}(A) &= \text{Rules}(A_1), \dots, \text{Rules}(A_n), \text{TopRule}(A). \end{aligned}$$

We only consider conflicts between arguments A and B such that the conclusion of A contradicts the conclusion of a subargument B' of B .

Conflicts between arguments having contradictory conclusions are resolved on the basis of a last-link ordering. An argument A is stronger than another argument B ($A > B$) if, and only if $\text{TopRule}(A)$ is stronger than $\text{TopRule}(B)$ [$\text{TopRule}(A) > \text{TopRule}(B)$]. Notice that we do not need to consider conflicts involving arguments of the form $\Rightarrow_F \phi$ since the set of facts is assumed to be consistent and no fact (or its negation) can occur in the head of any rule.³

Definition 5 (Defeats). An argument B *defeats* an argument A if, and only if $\exists A' \in \text{Sub}(A)$ such that $\text{Conc}(B) = \sim\text{Conc}(A')$, and $A' \not> B$.

An argument B *strictly defeats* an argument A if, and only if B defeats A and A does not defeat B .

We can now define the argumentation framework that is determined by an argumentation theory.

³ This simplification does not affect the generality of the approach. Such an assumption, which can be abandoned (see Governatori et al., 2004), allows for exploring some interesting properties of explanations: see Governatori et al. (2022b) and Section 5.4.

Definition 6. (Argumentation framework in structured argumentation). Let $D = (R, F, >)$ be an argumentation theory. The *argumentation framework* $AF(D)$ determined by D is (\mathcal{A}, \gg) where \mathcal{A} is the set of all arguments constructible from D , and \gg is the defeat relation defined above.

Given this definition of argumentation framework, if D is an argumentation theory, we can abuse notation somewhat and write $GE(D)$ to denote the grounded extension of the argumentation framework determined by D .

As noted above we consider that an argument is justified iff it is included in the grounded extension, and a conclusion justified iff it is supported by a justified argument.

Example 1. Consider the following theory D , describing a COVID scenario adapted from Italian temporary legal measures to prevent the spreading of pandemics (Governatori et al., 2022a).⁴

$$\begin{aligned} F &= \{positive, vax, \neg mask, old\} \\ R &= \{r_1 : positive, quarantine \Rightarrow \neg spread, \\ &\quad r_2 : positive \Rightarrow spread, \\ &\quad r_3 : positive, mask \Rightarrow \neg spread, \\ &\quad r_4 : spread, vax \Rightarrow \neg high_lethality, \\ &\quad r_5 : spread, old \Rightarrow high_lethality, \\ &\quad r_6 : high_lethality \Rightarrow hospital_collapse, \\ &\quad r_7 : positive \Rightarrow mask_obligatory, \\ &\quad r_8 : hospital_collapse \Rightarrow lockdown_obligatory\} \\ > &= \{\langle r_1, r_2 \rangle, \langle r_3, r_2 \rangle, \langle r_5, r_4 \rangle\}. \end{aligned}$$

Let us define the set \mathcal{A} of arguments from D :

$$\begin{aligned} \mathcal{A} &= \{A_1 : \Rightarrow_F positive, \\ &\quad A_2 : \Rightarrow_F vax, \\ &\quad A_3 : \Rightarrow_F \neg mask, \\ &\quad A_4 : \Rightarrow_F old, \\ &\quad A_5 : A_1 \Rightarrow_{r_2} spread, \\ &\quad A_6 : A_5, A_2 \Rightarrow_{r_4} \neg high_lethality, \\ &\quad A_7 : A_5, A_4 \Rightarrow_{r_5} high_lethality, \\ &\quad A_8 : A_7 \Rightarrow_{r_6} hospital_collapse, \\ &\quad A_9 : A_1 \Rightarrow_{r_7} mask_obligatory, \\ &\quad A_{10} : A_8 \Rightarrow_{r_8} lockdown_obligatory\}. \end{aligned}$$

The argumentation framework determined by D is thus $AF(D) = (\mathcal{A}, \gg)$ where

$$\gg = \{\langle A_7, A_6 \rangle\}.$$

The grounded extension of $AF(D)$ is $\{A_1, A_2, A_3, A_4, A_5, A_7, A_8, A_9, A_{10}\}$. The set $GE(D)$ of justified conclusions is

$$GE(D) = \{positive, vax, \neg mask, old, spread, high_lethality, hospital_collapse, mask_obligatory, lockdown_obligatory\}.$$

⁴ <https://www.osservatoriosullefonti.it/emergenza-covid-19/fonti-governative/decreti-del-presidente-del-consiglio-dei-ministri/2997-emcov-dpcm-elenco>

5. Types of explanation in legal argumentation

As informally discussed in Sections 2, 3, an open research issue concerns the relation between the *justification of arguments* and the *explanation of legal conclusions*. To address this issue, we shall try to build a bridge between two research lines using formal argumentation: the AI & Law investigation on the justification of (automated) legal decision-making, and the study of the idea of explanation in the context of eXplainable AI. The following sections provide some general ideas to fill the gap and aim at potentially addressing, at an abstract level, the challenges discussed in Sections 2, 3.

In rule-based systems, finding an explanation for a certain normative conclusion ϕ (such as a legal conclusion) requires determining if certain pieces of information support the conclusion of ϕ through a set of rules (Governatori et al., 2022b). In the context of argumentation, such an intuition should be adjusted and further elaborated. Notice that, in contrast with the majority of the literature (see Section 6) we provide several definitions of the idea of legal explanation that do not simply focus on arguments, but also that make an essential use of the distinctive elements (facts, rules, priorities) of argumentation frameworks.

Our contribution is mainly conceptual and it is meant to show how notions such as those proposed by Miller (2019) or discussed in legal theory can be modeled in an argumentation framework: an extensive formal study is left to future research.

5.1. Explanations by sufficient or necessary arguments

Let us first introduce two auxiliary notions, i.e., closure under subarguments and superarguments.

Definition 7. (Closure under subarguments and under superarguments). A set S of arguments is closed under subarguments iff for every arguments $A \in S$, $\text{Sub}(A) \subseteq S$.

A set S of arguments is closed under superarguments w.r.t. an argument set W , iff for every arguments $A \in W$ and $A' \in S$ such that $A' \in \text{Sub}(A)$, $A \in S$.

Let us begin with two basic concepts of legal explanation that draw inspiration from Hart and Honore's (1959) NESS theory of legal causation, and which are reframed here to cover arguments built using norms.

We start with the concept of *explanation by sufficient arguments*, by which we mean a minimal set of arguments which, within the given argumentation framework, is sufficient to determine a certain legal outcome.

Definition 8 (Explanation by sufficient arguments). Let $D = (R, F, >)$ be an argumentation theory and $AF(D) = (\mathcal{A}, \gg)$ be the argumentation framework determined by D . The set $\mathcal{E} \subseteq \mathcal{A}$ is an *explanation of ϕ by sufficient arguments w.r.t. D* iff

- $A \in \mathcal{E}$ is an argument for ϕ and A is justified w.r.t. D ;
- \mathcal{E} is a minimal set such that, for every argument $B \in \mathcal{A}$ that defeats A , there is an argument $C \in \mathcal{E}$ that strictly defeats B ;

- \mathcal{E} is closed under subarguments.

Notice that a broader concept of explanation by sufficient arguments for a conclusion ϕ could be obtained by the set-theoretical union of all explanations by sufficiency of ϕ .

Remark 1. The idea of explanation by sufficient arguments may be philosophically linked to [Hart and Honoré's \(1959\)](#) NESS approach to causality, where a cause for an effect is a necessary element of a sufficient set of conditions for that effect. In our framework, any explanation by sufficient arguments \mathcal{E} of ϕ is a sufficient set for ϕ .

Within formal argumentation, the idea of an explanation by sufficient arguments has been firstly elaborated with minor differences by [Fan and Toni \(2015\)](#) with the idea of related admissibility, which states that a set of arguments \mathcal{E} is relatedly admissible iff $\exists A \in \mathcal{E}$ s.t. \mathcal{E} defends A and \mathcal{E} is admissible. In particular, the authors identify a case where \mathcal{E} is minimal (they call this case *minimal explanation*). A difference with respect to our definition is that we focus on the conclusion ϕ (which can be supported by more than one argument) and not on a single argument. A similar analysis has been also proposed by [Borg and Bex \(2020\)](#).

The second notion of explanation of a proposition is that of *explanation by necessary arguments*. This includes a set of arguments such that their omission from the argumentation framework would prevent the proposition being justified. Note that this notion is independent from the notion of explanation by sufficient arguments, as introduced in Definition 8.

Definition 9 (Explanation by necessary arguments).

Let $D = (R, F, >)$ be an argumentation theory and $AF(D) = (\mathcal{A}, \gg)$ be the argumentation framework determined by D , and ϕ be a justified conclusion of $AF(D)$. The set $\mathcal{E} \subseteq \mathcal{A}$ is an *explanation by necessary arguments* of ϕ w.r.t. $AF(D)$ iff

- ϕ is not justified w.r.t. $AF'(D') = (\mathcal{A}/S, \gg')$, where S is the closure under superarguments of \mathcal{E} relatively to \mathcal{A} and $\gg' = \gg - \{(A, B) \mid A \in S \text{ or } B \in S\}$;
- \mathcal{E} is minimal.

Example 2. According to Definition 9, assume that $AF(D)$ contains argument $[[a] \Rightarrow b] \Rightarrow c$ as well as argument $[d] \Rightarrow c$. Then c is explained through necessary arguments by any set including a subargument for each of these arguments. For instance c is explained by $\{[a] \Rightarrow b, d\}$, since c cannot be established if both $[a] \Rightarrow b$ and d were not available.

Remark 2. Notice that [Borg and Bex \(2020\)](#) have also considered the explanation by necessary arguments. In this work, however, the focus X is on single arguments and the target (i.e., the Y for which X is necessary) is an argument and not a legal conclusion ϕ (a conclusion can in fact be supported by more arguments). For this reason, the authors do not explicitly state that, when considered more necessary arguments, S must be closed under superarguments.

In legal reasoning often the rules are assumed to be fixed and we only consider the facts as relevant explanations. For instance, if

asked why one got a fine, a sufficient answer may consist in pointing to the fact that the speed was 100 km per hour, if it is fixed the set of norms containing the rule prohibiting such a speed.

Following this idea, we can provide the following notions of explanations by sufficient and necessary facts, extracting factual arguments from explanations by sufficient and necessary arguments.

Definition 10 (Explanation by sufficient/necessary facts).

Let $D = (R, F, >)$ be an argumentation theory and $AF(D) = (\mathcal{A}, \gg)$ be the argumentation framework determined by D . The set \mathcal{F} is an *explanation of ϕ by sufficient/necessary facts w.r.t. D* iff

- \mathcal{E} is an explanation by sufficient/necessary arguments of ϕ and
- \mathcal{F} is the set of all and only the factual arguments in \mathcal{E} .

5.2. Contrastive explanations

Let us now consider some specifications of an idea of explanation that is well-known in the literature ([Miller, 2019](#)), which is widely used in XAI ([Miller et al., 2022](#)), and which has been recently considered in the context of legal reasoning ([Borg and Bex, 2020](#); [Liu et al., 2022a](#)). We may informally characterize such explanations as follows:

Intuition 1 (Contrastive explanation). Saying that ϕ is contrastively explained by x' means saying that if x' rather than x had been the case, then ϕ' rather than ϕ would have been the case.

We may develop the intuition above depending on whether we consider facts or rules. Indeed, the idea for modeling such a notion is to remove/add relevant facts or rules in such a way that the justification status of ϕ will change, and use these changes to provide (part of) an explanation (see [Liu et al., 2022b](#), following [Miller, 2019](#)).

Note that our notion of a contrastive explanation covers two different ways in which the justification of a proposition can be interfered with. The interference may consist in (a) removing from the theory elements being used in arguments that directly or indirectly support the proposition at stake or (b) inserting in the theory elements to be used in arguments that directly or indirectly attack the proposition at stake. Obviously, indirect support consists in attacking attackers and indirect attack in attacking defenders.

Let us first focus on the facts (the literals) that are being used to build legal arguments. We then consider what arguments would be available if the set of facts were changed, adding and/or removing some facts. Thus the contrastive explanation is obtained by considering a minimal pair $\langle F^-, F^+ \rangle$ where F^- are the facts to be deleted, and F^+ the facts to be consistently added (i.e., such that $F \cup F^+$ is consistent).

Definition 11 (Fact-based contrastive explanation). Let $D = (R, F, >)$ be an argumentation theory and ϕ be justified w.r.t. D . Then $\langle F^-, F^+ \rangle$ is a *fact-based contrastive explanation of ϕ w.r.t. $AF(D)$* iff

1. $(F \setminus F^-) \cup F^+$ is consistent;

2. ϕ is not justified w.r.t. $D' = (R, (F \setminus F^-) \cup F^+), >);$
3. no $\langle F'^-, F'^+, \rangle$, where $F'^- \cup F'^+ \subset F^+ \cup F^-$, satisfies conditions 1 and 2.

Example 3. Let us apply Definition 11 to Example 1 above. It appears that a fact-based contrastive explanation for *lockdown_obligatory* is provided by $\langle \{positive\}, \emptyset \rangle$: *positive* contrastively explains that outcome since, without this fact the explanandum would not be justified (if positivity were not the case there would be no obligatory lockdown). Another explanation for the same explanandum would be $\langle \{old\}, \emptyset \rangle$.

Similarly, $\langle \{\neg mask\}, \{mask\} \rangle$ is an explanation for *lockdown_obligatory*, since if people had masks rather than not having them, the explanandum would not hold. In fact, under such a change, all the rest remaining the same, we can infer $\neg spread$ so defeating the argument for spread. This would prevent the derivation of *high_letality*, *hospital_collapse* and *lockdown_obligatory*.

Besides contrastively explaining a proposition ϕ , as in Definition 11, we may also contrastively explain the non-acceptance of a proposition relative to a theory, i.e., of the failure to provide a justification for it.

Definition 12. (Fact-based contrastive explanation of non-acceptance). Let $D = (R, F, >)$ be an argumentation theory and ϕ not be justified w.r.t. D . Then $\langle F^-, F^+ \rangle$, is a *fact-based contrastive explanation of the non-acceptance of ϕ w.r.t. $AF(D)$* iff

1. $(F \setminus F^-) \cup F^+$ is consistent;
2. ϕ is justified w.r.t. $D' = (R, (F \setminus F^-) \cup F^+), >);$
3. no $\langle F'^-, F'^+ \subseteq F^+ \rangle$, where $F'^- \cup F'^+ \subset F^+ \cup F^-$, satisfies conditions 1 and 2.

Example 4. Consider again Example 1 add to it the following rule, according which if the pandemic does not spread, we can have a normal life under the pandemic:

$$r_9 : \neg spread \Rightarrow normal_life$$

We may than ask “Why is it that we cannot have a normal life,” and an answer would be the contrastive explanation $\langle \{\neg mask\}, \{mask\} \rangle$: people are not wearing masks (rather than wearing them). In fact, after the theory is revised by removing, $\neg mask$ and adding *mask*, there is a justified argument for *normal_life*, based on rule r_9 , whose antecedent condition $\neg spread$ can be establishes by using rule r_3 , and facts *positive*, and *mask*.

The ideas just described can be expanded by assuming that also rules can be removed or added. The rules to be removed are included in the current theory, while the rules to be added can be built from the language (see Definition 2). Thus we obtain the following definition, which matches Definition 11 above.

Definition 13 (Rule-based contrastive explanation). Let $D = (R, F, >)$ be an argumentation theory and ϕ be justified w.r.t. D . Then $\langle R^-, R^+ \rangle$, with $R^- \subseteq R$ and $R^+ \subseteq \text{Rul}$, is a *rule-based contrastive explanation of ϕ w.r.t. $AF(D)$* iff

1. $D' = (R \setminus R^-, R^+, F, >')$ where $>' = > - \{(r, r') \mid \{r, r'\} \cap R^- \neq \emptyset\}$;

2. ϕ is not justified w.r.t. D' ;
3. no $\langle R'^-, R'^+ \rangle$, such that $(R'^- \cup R'^+) \subset (R^- \cup R^+)$, satisfies conditions 1 and 2.

Finally, by combining the possibility to add or remove facts, rules, or even rule-priorities, we come to the following definition:

Definition 14. (Fact-, rule-, and priority-based contrastive explanation). Let $D = (R, F, >)$ be an argumentation theory, $AF(D) = (\mathcal{A}, \gg)$ be the argumentation framework determined by D , and ϕ be justified w.r.t. D . Then $\langle F^-, F^+ \rangle, \langle R^-, R^+ \rangle, \langle >^-, >^+ \rangle$, with $F^-, F^+ \subseteq \text{ANT}(R)$, $R^-, R^+ \subseteq \text{Rul}$, $>^-, >^+ \subseteq \text{Rul} \times \text{Rul}$ is a *fact-rule-priority-based contrastive explanation of ϕ w.r.t. $AF(D)$* iff

1. $AF(D') = (\mathcal{A}, \gg)$ is the argumentation framework determined by $D' = (R \setminus R^-, R^+, F \setminus F^-) \cup F^+, (> \setminus >^-) \cup >^+)$
2. ϕ is not justified wrt D' ;
3. Conditions 1 and 2. are satisfied by no triplet $\langle F'^-, F'^+ \rangle, \langle R'^-, R'^+ \rangle, \langle >'^-, >'^+ \rangle$, such that $\cup(F'^-, F'^+, R'^-, >'^-, >'^+) \subset \cup(F^-, F^+, R^-, >^-, >^+)$.

The definitions above are abstract and fit the structure of argumentation frameworks: the effective process of defining minimal revisions of rules and priorities is rather complex (see Billington et al., 1999; Governatori and Rotolo, 2010; Boella et al., 2016; Governatori et al., 2019).

Example 5. Consider again Example 1 and the normative conclusion *lockdown_obligatory*. Trivially, $\langle \{r_2\}, \emptyset \rangle, \langle \{r_5\}, \emptyset \rangle, \langle \{r_6\}, \emptyset \rangle$, and $\langle \{r_8\}, \emptyset \rangle$ are rule-based contrastive explanations of *lockdown_obligatory* w.r.t. $AF(D)$.

Assume that F would already include the fact *countryside* and suppose to change the theory D into $D' = (R', F, >)$ as follows:

$$R' = R \cup \{r_9 : countryside, spread \Rightarrow \neg lockdown_obligatory\}.$$

Then, we would have two new arguments

$$A_{12} : \Rightarrow_F countryside,$$

$$A_{13} : A_5, A_{12} \Rightarrow_{r_9} \neg lockdown_obligatory$$

Since we work in a skeptical semantics, $\langle \{r_9\}, \emptyset \rangle$ is a rule-based contrastive explanation of *lockdown_obligatory* w.r.t. $AF(D)$.

Finally, suppose we obtain D' by simply making $>$ empty: then, $\langle \emptyset, \emptyset \rangle, \langle \emptyset, \emptyset, >^-, \emptyset \rangle$ is a fact-rule-priority-based contrastive explanation of *lockdown_obligatory* w.r.t. $AF(D)$.

5.3. Discussion and further examples

Contrastive explanation is perhaps the best example to highlight the third-party nature of explanations as discussed in Section 2. Indeed, such a type of explanation explicitly compares two different argumentation theories and frameworks, which could in fact correspond to two different argumentative angles: one could be attributed to the decision-maker and one of to any observer that rationally reconstructs the decision and explains it by comparison.

More precisely, the actual argumentation framework where we justify a certain legal conclusion provides the perspective of the

decision-maker \mathcal{D} , while the comparison between this framework and anything else is made by a neutral observer \mathcal{O} .

Example 6. Let us go back to the case of legal interpretation briefly recalled in Section 3.4 and consider the following provision from the Italian penal code:

Art. 575. Homicide. Whoever causes the death of a *man* [*uomo*] is punishable by no less than 21 years in prison.

The almost unanimous interpretation of courts of art. 575 is that, of course, it covers killing of *any* person and not only of men. For doing so, one may consider the ordinary interpretation of art. 3 of the Italian constitution, which establishes, among other things, that all people have equal social status and are equal before the law, without regard to any personal aspects including gender. This is an argument from general principles. Alternatively, one may use an argument by coherence and maintain that the ordinary interpretation of other legislative provisions n does the same. Both exclude the ordinary reading of “man” as “adult male human being.” Consider the following argumentation theory D , where ψ means “only the death of a human male is punishable by no less than 21 years of prison”:

$$\begin{aligned} R = \{ & r' : \text{I}_{\text{ordinary}}(\text{art.3}, \pi) \Rightarrow \text{I}_{\text{constitutional_principle}}(\text{art.575}, \neg\psi) \\ & s' : \text{I}_{\text{ordinary}}(n, \gamma) \Rightarrow \text{I}_{\text{coherence}}(\text{art.575}, \neg\psi), \\ & t' : \Rightarrow \text{I}_{\text{ordinary}}(\text{art.575}, \psi) \\ & z' : \text{I}_{\text{constitutional_principle}}(\text{art.575}, \neg\psi) \Rightarrow \neg\psi, \\ & z'' : \text{I}_{\text{coherence}}(\text{art.575}, \neg\psi) \Rightarrow \neg\psi, \\ & z''' : \text{I}_{\text{ordinary}}(\text{art.575}, \psi) \Rightarrow \psi \} \\ F = \{ & \text{I}_{\text{ordinary}}(\text{art.3}, \pi), \text{I}_{\text{ordinary}}(n, \gamma) \} \\ \geq = \{ & \langle z', z''' \rangle \}. \end{aligned}$$

Suppose a court decides a case rejecting ψ and supports $\neg\psi$ because of r' , i.e., in the light of art. 3. Indeed, since \mathcal{A} in $\text{AF}(D)$ includes

$$\begin{aligned} A_1 : & \Rightarrow_F \text{I}_{\text{ordinary}}(\text{art.3}, \pi) \\ A_2 : & A_1 \Rightarrow_{r'} \text{I}_{\text{constitutional_principle}}(\text{art.575}, \neg\psi) \\ A_3 : & A_2 \Rightarrow_{z'} \neg\psi \end{aligned}$$

then the argument A_3 and its conclusion $\neg\psi$ using r' and z' are justified in the corresponding argumentation framework $\text{AF}(D)$.

Preliminarily, we should note that

- $\{A_3, A_2, A_1\}$ is an explanation by sufficient arguments of $\neg\psi$;
- $\{A_3, A_2, A_1\}$ is not an explanation by necessary arguments of $\neg\psi$ if we added the rules

$$\begin{aligned} w : & \text{I}_{\text{principle}}(\text{art.575}, \neg\psi) \Rightarrow \text{I}_{\text{teleological}}(\text{art.575}, \neg\psi), \\ w' : & \text{I}_{\text{teleological}}(\text{art.575}, \neg\psi) \Rightarrow \neg\psi, \end{aligned}$$

and changed the priorities as follows

$$\geq = \{ \langle z', z''' \rangle, \langle w', z''' \rangle \}$$

being still $\{A_3, A_2, A_1\}$ an explanation by sufficiency.

This could be enough in the perspective of the decision-maker \mathcal{D} . Let us rationally reconstruct \mathcal{D} 's decision. Such a reconstruction may correspond to an observer \mathcal{O} : several options are available. Let us see three of them for the sake of illustration.

1. If $F^- = \{\text{I}_{\text{ordinary}}(\text{art.3}, \pi)\}$ then $\langle \emptyset, F^- \rangle$ is a fact-based contrastive explanation of ψ : \mathcal{O} 's explanation of \mathcal{D} 's decision in favor of ψ is based on noticing that this fact, if removed, would prevent the conclusion.
2. Since \mathcal{A} in $\text{AF}(D)$ includes the following set of justified arguments

$$\begin{aligned} A_1 : & \Rightarrow_F \text{I}_{\text{ordinary}}(\text{art.3}, \pi) \\ A_2 : & A_1 \Rightarrow_{r'} \text{I}_{\text{constitutional_principle}}(\text{art.575}, \neg\psi) \\ A_3 : & A_2 \Rightarrow_{z'} \neg\psi \\ A_4 : & \Rightarrow_F \text{I}_{\text{ordinary}}(n, \gamma) \\ A_5 : & A_4 \Rightarrow_{s'} \text{I}_{\text{coherence}}(\text{art.575}, \neg\psi) \\ A_6 : & A_5 \Rightarrow_{z''} \neg\psi \end{aligned}$$

while \mathcal{D} could only explicitly rely on A_3 , the observer \mathcal{O} would contrastively explain the decision by noticing that $\langle \{r', s'\}, \emptyset \rangle$ is rule-based contrastive explanation of $\neg\psi$.

3. Finally, assume to change the argumentation theory in such a way that $\geq = \emptyset$. Then \mathcal{D} would not decide in favor of $\neg\psi$. Since we work in skeptical argumentation, an observer \mathcal{O} can explain this decision by identifying elements that would be needed to conclude $\neg\psi$ and by simply noting that

$$\langle \emptyset, \emptyset \rangle, \langle \{r', s'\}, \emptyset \rangle, \langle \emptyset, \{z', z'''\} \rangle$$

is a fact-rule-priority-based contrastive explanation of $\neg\psi$.

5.4. Stable argumentative explanations

An interesting issue for investigation is the concept of *stable explanation* in argumentation, a concept that was explored from a proof-theoretic perspective, among others, by Brewka et al. (2019); Brewka and Ulbricht (2019); Governatori et al. (2022b). In particular, Governatori et al. (2022b,c) considered the problem of determining a stable normative explanation for a certain legal conclusion, which means to identify a set of facts (i.e., reasoning inputs) able to ensure that such a conclusion continues to hold when new facts are added to a normative case. The basic intuition is the following.

Intuition 2 (Stable explanation). A normative explanation for a given legal conclusion ϕ is stable when adding new normative elements to that explanation does not affect its power to explain ϕ .

Interestingly, in the context of legal argumentation, we can observe the following (Governatori et al., 2022c):

- Given the facts of the normative case, any judicial proceeding has the objective of determining what *legal requirements* (e.g., obligations, prohibitions, permissions, ascription of rights) hold, and whether such legal requirements have been fulfilled;

- If new facts were presented by one party in the proceeding, the outcome of the case could change;
- Each party in the judicial proceeding is thus interested in the following question: *How to ensure a specific outcome for a case, which, in an adversarial context, means how to ensure that the facts presented by such a party are “resilient” to the attacks of the opponent?*

The following example is adapted from Australian commercial law and from Governatori et al. (2022b,c), and illustrates the idea.⁵

Example 7. Suppose the law forbids private individuals engaging in credit activities. However, such activities are permitted if you have a credit license. Moreover, they are also permitted if you are acting on behalf of another person (the principal), who holds a credit license. In any case, such activities are prohibited if you have been banned from them by the competent regulatory authority. Consider the following theory D :

$$\begin{aligned} F &= \emptyset \\ R &= \{s_1 : \text{creditActivity} \Rightarrow \text{violation}, \\ &\quad s_2 : \text{creditLicense}, \text{creditActivity} \Rightarrow \neg \text{violation}, \\ &\quad s_3 : \text{actsOnBehalfPrincipal}, \text{principalCreditLicense}, \\ &\quad \text{creditActivity} \Rightarrow \neg \text{violation}, \\ &\quad s_4 : \text{banned}, \text{creditActivity} \Rightarrow \text{violation}\} \\ &\geq = \{\langle s_2 > s_1 \rangle, \langle s_3 > s_1 \rangle, \langle s_4 > s_2 \rangle, \langle s_4 > s_3 \rangle\}. \end{aligned}$$

It is easy to see that relative to theory D we can distinguish stable and unstable explanations:

- $F \cup \{\text{creditActivity}\}$ is not a stable explanation for *violation* w.r.t. D , since it is no explanation for *violation* in D' if facts are $\{\text{creditActivity}, \text{creditLicense}\}$ (*violation* not being a justified conclusion w.r.t. D');
- $F \cup \{\text{banned}, \text{creditActivity}\}$ is stable explanation for *violation* w.r.t. D , since there are no facts F' consistent with F (and with the conclusions of the rules in R) such that F is not an explanation of *violation* with regard to $D' = (R, F \cup F', >)$.

Here is a definition of a stable normative explanation, based on the analysis just provided. In the context of stable explanation by sufficient facts we need to consider facts that (a) are additional to the facts in the theory (b) are consistent with the such facts.

Definition 15 (Stable explanation by sufficient facts). Let $D = (R, F, >)$ be an argumentation theory and \mathcal{F} be the set of factual arguments of $\text{AF}(D)$. An explanation $\mathcal{E}' \subseteq \mathcal{F}$ by sufficient facts is *stable relative to D* if there is no set of facts F' such that

- $F \cap F' = \emptyset$,
- F' is consistent with F , and
- \mathcal{E}' is not an explanation by sufficient facts relative to $D' = (R, F \cup F', >)$.

It is easy to check that this definition works relative to the examples above. For instance, $\mathcal{E}' = \{\Rightarrow_F \text{creditActivity}\}$ is no

stable explanation by sufficient facts of *violation*, since adding $\{\text{creditLicense}\}$ to the facts is such that there is no explanation of *violation* relative to the facts $\{\text{creditActivity}, \text{creditLicense}\}$.

A broader account of Governatori et al. (2022b)'s approach is rule-based and proof-theoretic (in Defeasible Logic: Antoniou et al., 2001) while a deontic extension of it has been developed by Governatori et al. (2022c) to characterize the idea of deontic explanation. Relative to an argumentation setting such as the one from Section 4, we can establish the following theorem (for the proof, see Appendix).

Theorem 1. Given a theory D and an explanation by sufficient facts \mathcal{F} relative to D , the problem of determining if \mathcal{F} is stable is co-NP-complete.

6. Related and future work

We have provided multiple characterisations for the idea of normative explanation in legal argumentation. We hope that our work, though coherent with previous literature, may contribute to further developments on the interaction between argumentation and explanation in the legal domain. The following lines of inquiry are especially relevant to our endeavor:

- Research on explanation in argumentation;
- Research on explanation in the AI & Law domain;
- Research on norm revision and other issues in legal reasoning.

6.1. Explanation in argumentation

The idea of modeling explanations in an argumentation framework for decision-making is not new (for an overview, see Cyras et al., 2021b). Approaches to argument-based decision-making have been developed, where argumentation is used to evaluate arguments for and against potential decisions, with the argumentation frameworks constituting the explanations (Amgoud and Prade, 2009). Our approach is connected to this idea, though we extract explanations from argumentation frameworks, rather than viewing argumentation framework as explanations.

The goal of providing explanation through argumentation has inspired the research by Toni et al. starting from (Fan and Toni, 2015) [several subsequent contributions appeared and recent developments have been proposed in several applied fields such as medical diagnostics (Cyras et al., 2021a)]. They construct arguments using rules as we do and elaborate the idea of explanation in an argument-based way (also considered in Cyras et al., 2021b). They see *explanation of an argument A* as a relation between A and a subset \mathcal{E} of a set of admissible set of arguments to which A belongs. Different appropriateness criteria are adopted to define \mathcal{E} , according to which explanations can be classified into different types: minimal explanation, compact explanation, maximal explanation, and so forth. Differently from them we have focused the need to provide an appropriate *explanation for a legal conclusion*, i.e., and explanation that may be meaningful for the humans involved (relying on Miller, 2019), thus focusing particularly on contrastive explanations. Our work is also related

⁵ <https://www.legislation.gov.au/Details/C2009A00134>

to Borg and Bex (2021a,b), who propose similar definitions of explanation by sufficient and necessary arguments, but who do not consider several contrastive models.

Other relevant contributions in decision-making are Liao and van der Torre (2020) and Besnard et al. (2022), which however reconstruct explanations within an abstract argumentation perspective.

6.2. Explanation in legal argumentation and AI & Law

The concept of explanation has played an important role in the AI & Law community, being related with the general quest of justification and transparency of legal decision-making (Atkinson et al., 2020). Within this community, argument-based explanations have been considered in the domain of evidence (Walton, 2005; Di Bello and Verheij, 2020), as well as in case-based reasoning (Liu et al., 2022a; Prakken and Ratsma, 2022). Prakken and Ratsma (2022) reconstruct explanations—and in particular contrastive explanations in the context as argument games between a proponent and opponent of an argument (i.e., a case citation for an outcome to be explained). Liu et al. (2022a) directly follow Miller (2019) and argue that a case base can be represented through a binary classifier: thus contrastive and counterfactual explanations are used to explain the outcomes of the classifier. Though valuable, those systems work on cases having the form $c = (s, X, c)$, where s is a state/fact situation, $c \in \{0, 1\}$ (the outcome favors the defendant or the plaintiff), and X , called the *reason* of the decision, is a subset of s . The structure of decisions and legal reasoning is much richer in our framework.

An interesting contribution in legal reasoning—but mainly focused on legal evidence—is Borg and Bex (2020), which develops similar notions of explanation by sufficient and necessary arguments. The idea of contrastive is also considered, but the approach is technically rather different. The authors, given the question “why P rather than Q ?”, call P the fact and Q the foil (Lipton, 1990). The contrastive explanation aims at making the foil explicit and considers those arguments that explain: (a) the acceptance of the fact and the non-acceptance of the foil; (b) the non-acceptance of the fact and the acceptance of the foil. Our approach provide several options that exploit the structure of argumentation theories, and which are not discussed by Borg and Bex (2020) (such as the distinction between factual and plain arguments).

6.3. Norm revisions and legal reasoning

As we have shown, the idea of contrastive and stable explanation require the current argumentation framework to be changed. Hence, an interesting issue is rethinking the quest for an explanation as an abductive inference, based on the revision of the given argumentation theory (Governatori and Rotolo, 2010; Governatori et al., 2019). Formally, given the argumentation theory D_{init} , the revised theory D , and the

target conclusion ϕ , we could formally define change operations as follows:

Expansion: from $D_{init} \not\vdash \phi$ to $D \vdash \phi$.

Contraction: from $D_{init} \vdash \phi$ to $D \not\vdash \phi$.

Revision: from $D_{init} \vdash \phi$ to $D \vdash \sim\phi$.

The development of this intuition has to be left to future research.

Another interesting future development concerns the import of the proposed idea of explanation in legal theory. While it is well-known that the idea of explanation can be used to reconstruct causality, it is less clear how to apply it to normative reasons. It can be interesting to mention here an exponent of classical doctrine of case law, Wambaugh (1894), who stated that the identification of the ratio decidendi of a precedent starting from a particular datum—understood as part of the argumentative framework—is reduced to a procedure in which one must ask whether, by denying this datum, the court could reach the conclusion obtained. This suggests that various types of explanation can play an interesting role in case-based reasoning (Liu et al., 2022a), including the idea of counterfactual explanation (Miller, 2019), which is left as well to future research.

7. Conclusion

In this paper we have discussed the role of argumentation in the law, and reviewed some literature of formal models of legal argumentation. Then we have investigated the formal connection between argumentation and explanation in the law. In particular, we have proposed several definitions of an explanation in the context of formal argumentation, articulating the relations between the justification of arguments and explanations.

One basic theoretical challenge was at the core of our contribution: clarifying through formal argumentation the structure in normative reasoning of the concepts of justification and explanation. In legal theory, the focus usually is on providing a justification for legal decisions, so that the idea of an explanation only plays a secondary role. This is due to the fact that on the one hand it is assumed that legal decision-making requires strong standard of (internal) rationality, and on the other hand the notion of an explanation is usually confined to what we called causal explanation, rather than to rational reconstruction.

In this paper we took a different perspective, which is closer to how the concept of explanation has been formally developed in logic and adopted in XAI. We argued that the distinction between justification and explanation is pragmatical rather than structural. Thus we can include rational reconstructions within the scope of explanation, and have argued that such reconstructions can be extracted from justifications, to provide an account of the logic of such justification with regard to the issues at stake. Thus, we have developed various notions of explanation on top of the justification of arguments and conclusions, such as different kinds of contrastive explanations.

We have also presented the idea of stable normative explanation (Governatori et al., 2022c). The problem of determining a stable normative explanation for a certain legal conclusion means to identify a set of facts, obligations, permissions, and other normative inputs able to ensure that such a conclusion continues to hold when new facts are added to a case. This notion is interesting from a logical point of view—think about the classical idea of inference to the best explanation—but it can contribute to symbolic models for XAI for the law (consider, for instance, systems of predictive justice).

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

All authors contributed to the article and approved the submitted version.

References

- Achinstein, P. (1983). *The Nature of Explanation*. Oxford: Oxford University Press.
- Alchourron, C., and Bulygin, E. (1971). *Normative Systems. LEP Library of Exact Philosophy*. Vienna: Springer. doi: 10.1007/978-3-7091-7118-9
- Aleven, V., and Ashley, K. D. (1997). "Evaluating a learning environment for case-based argumentation skills," in *Proceedings of the Sixth International Conference on Artificial Intelligence and Law (ICAIL-97)* (New York, NY: ACM), 170–179. doi: 10.1145/261618.261650
- Alexy, R. (1989). *A Theory of Legal Argumentation: The Theory of Rational Discourse as Theory of Legal Justification*. Oxford: Clarendon.
- Amgoud, L., and Prade, H. (2009). Using arguments for making and explaining decisions. *Artif. Intell.* 173, 413–436. doi: 10.1016/j.artint.2008.11.006
- Antoniou, G., Billington, D., Governatori, G., and Maher, M. J. (2001). Representation results for defeasible logic. *ACM Trans. Comput. Log.* 2, 255–286. doi: 10.1145/371316.371517
- Ashley, K. D., and Aleven, V. (1991). "Toward and intelligent tutoring system for teaching law students to argue with cases," in *Proceedings of the Third International Conference on Artificial Intelligence and Law (ICAIL)* (ACM), 42–52. doi: 10.1145/112646.112651
- Ashley, K. D. (1990). *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*. Cambridge, MA: MIT.
- Atkinson, K., Bench-Capon, T., and Bollegala, D. (2020). Explanation in AI & Law: past, present and future. *Artif. Intell.* 289, 103387. doi: 10.1016/j.artint.2020.103387
- Atkinson, K., and Bench-Capon, T. J. M. (2007). Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artif. Intell.* 171, 855–874. doi: 10.1016/j.artint.2007.04.009
- Baier, K. (1958). *The Moral Point of View*. Ithaca, NY: Cornell University Press.
- Baroni, P., and Giacomini, M. (2009). "Semantics of abstract argument systems," in *Argumentation in Artificial Intelligence* (Berlin, Heidelberg: Springer), 25–44. doi: 10.1007/978-0-387-98197-0_2
- Bench-Capon, T. J. M., and Rissland, E. L. (2002). A note on dimensions and factors. *Artif. Intell. Law* 10, 65–77. doi: 10.1023/A:1019501830692
- Bench-Capon, T. J. M., and Sartor, G. (2003). A model of legal reasoning with cases incorporating theories and values. *Artif. Intell.* 150, 97–142. doi: 10.1016/S0004-3702(03)00108-5
- Bench-Capon, T., Prakken, H., Wyner, A., and Atkinson, K. (2013). "Argument schemes for reasoning with legal cases using values," in *Proceedings of the 14th International Conference on Artificial Intelligence and Law* (New York, NY: ACM), 13–22. doi: 10.1145/2514601.2514604
- Berman, D. H., and Hafner, C. D. (1993). "Representing teleological structure in case-based reasoning: the missing link," in *Proceedings of the Fourth International Conference on Artificial Intelligence and Law (ICAIL)* (New York, NY: ACM), 50–59. doi: 10.1145/158976.158982
- Besnard, P., Doutre, S., Duchatelle, T., and Lagasque-Schiex, M.-C. (2022). Explaining semantics and extension membership in abstract argumentation. *Intell. Syst. Appl.* 16, 200118. doi: 10.1016/j.iswa.2022.200118
- Billington, D., Antoniou, G., Governatori, G., and Maher, M. (1999). Revising nonmonotonic theories: the case of defeasible logic. *Lect. Notes Comput. Sci.* 1701, 101–112. doi: 10.1007/3-540-48238-5_8
- Boella, G., Pigozzi, G., and van der Torre, L. (2016). Agm contraction and revision of rules. *J. Log. Lang. Inf.* 25, 273–297. doi: 10.1007/s10849-016-9244-9
- Bongiovanni, G., Postema, G., Rotolo, A., Sartor, G., Valentini, C., Walton, D., et al. (2018). *Handbook of Legal Reasoning and Argumentation*. Berlin: Springer. doi: 10.1007/978-90-481-9452-0
- Borg, A., and Bex, F. (2021a). A basic framework for explanations in argumentation. *IEEE Intell. Syst.* 36, 25–35. doi: 10.1109/MIS.2021.3053102
- Borg, A., and Bex, F. (2020). "Explaining arguments at the dutch national police," in *AI Approaches to the Complexity of Legal Systems XI-XII - AICOL International Workshops 2018 and 2020: AICOL-XI@JURIX 2018, AICOL-XII@JURIX 2020, XAILA@JURIX 2020, Revised Selected Papers, Volume 13048 of Lecture Notes in Computer Science*, eds V. Rodríguez-Doncel, M. Palmirani, M. Araszkievicz, P. Casanovas, U. Pagallo, and G. Sartor (Berlin: Springer), 183–197. doi: 10.1007/978-3-030-89811-3_13
- Borg, A., and Bex, F. (2021b). "Necessary and sufficient explanations for argumentation-based conclusions," in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 16th European Conference, ECSQARU 2021, Prague, Czech Republic, September 21-24, 2021. Proceedings, volume 12897 of Lecture Notes in Computer Science*, eds J. Vejnárová, and N. Wilson (Berlin: Springer), 45–58. doi: 10.1007/978-3-030-86772-0_4
- Brewka, G., Thimm, M., and Ulbricht, M. (2019). Strong inconsistency. *Artif. Intell.* 267, 78–117. doi: 10.1016/j.artint.2018.11.002
- Brewka, G., and Ulbricht, M. (2019). "Strong explanations for nonmonotonic reasoning," in *Description Logic, Theory Combination, and All That, Volume 11560 of LNCS* (Berlin: Springer), 135–146. doi: 10.1007/978-3-030-22102-7_6

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. 833647).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Calegari, R., and Sartor, G. (2021). "Burdens of persuasion and standards of proof in structured argumentation," in *Logic and Argumentation 4th International Conference, CLAR 2021 Hangzhou, China, October 20-22, 2021. Proceedings*, eds P. Baroni, C. Benz Müller, and Y. N. Wang (Berlin: Springer), 40–459. doi: 10.1007/978-3-030-89391-0_3
- Cyras, K., Oliveira, T., Karamlou, M., and Toni, F. (2021a). Assumption-based argumentation with preferences and goals for patient-centric reasoning with interacting clinical guidelines. *Argum. Comput.* 12, 149–189. doi: 10.3233/AAC-200523
- Cyras, K., Rago, A., Albini, E., Baroni, P., and Toni, F. (2021b). "Argumentative XAI: a survey," in *Proc. IJCAI-2021*. doi: 10.24963/ijcai.2021/600
- da Costa Pereira, C., Liao, B., Malerba, A., Rotolo, A., Tettamanzi, A. G. B., van der Torre, L. W. N., et al. (2017). Handling norms in multi-agent systems by means of formal argumentation. *FLAP 4*, 3039–3073.
- Davidson, D. (1963). Actions, reasons, and causes. *J. Philos.* 60, 685. doi: 10.2307/2023177
- Di Bello, M., and Verheij, B. (2020). Evidence and decision making in the law: theoretical, computational and empirical approaches. *Artif. Intell. Law* 28, 1–5. doi: 10.1007/s10506-019-09253-0
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77, 321–358. doi: 10.1016/0004-3702(94)00041-X
- Fan, X., and Toni, F. (2015). On computing explanations in argumentation. *Proc. AAAI Conf. Artif. Intell.* 29, 1496–1502. doi: 10.1609/aaai.v29i1.9420
- Gordon, T. F. (1995). *The Pleadings Game. An Artificial Intelligence Model of Procedural Justice*. Dordrecht: Kluwer. doi: 10.1007/978-94-015-8447-0
- Governatori, G., and Rotolo, A. (2023). "Deontic ambiguities in legal reasoning," in *ICAIL 2023* (New York, NY: ACM).
- Governatori, G., and Rotolo, A. (2010). Changing legal systems: legal abrogations and annulments in defeasible logic. *Log. J. IGPL* 18, 157–194. doi: 10.1093/jigpal/jzp075
- Governatori, G., Maher, M. J., Antoniou, G., and Billington, D. (2004). Argumentation semantics for defeasible logics. *J. Log. Comput.* 14, 675–702. doi: 10.1093/logcom/14.5.675
- Governatori, G., Olivieri, F., Cristani, M., and Scannapieco, S. (2019). Revision of defeasible preferences. *Int. J. Approx. Reason* 104, 205–230. doi: 10.1016/j.ijar.2018.10.020
- Governatori, G., Olivieri, F., Rotolo, A., and Cristani, M. (2022a). "From defeasible logic to counterfactual reasoning," in *Proc. Declarative AI 2022* (Cham: Springer).
- Governatori, G., Olivieri, F., Rotolo, A., and Cristani, M. (2022b). "Inference to the stable explanations," in *LPNMR 2022* (Cham: Springer), 245–258. doi: 10.1007/978-3-031-15707-3_19
- Governatori, G., Olivieri, F., Rotolo, A., and Cristani, M. (2022c). "Stable normative explanations," in *Legal Knowledge and Information Systems - JURIX 2022: The Thirty-fifth Annual Conference, Saarbrücken, Germany, 14-16 December 2022, volume 362 of Frontiers in Artificial Intelligence and Applications*, eds E. Francesconi, G. Borges, and C. Sorge (Amsterdam: IOS Press), 43–52. doi: 10.3233/FAIA220447
- Governatori, G., Olivieri, F., Rotolo, A., Scannapieco, S., and Sartor, G. (2014). "Two faces of strategic argumentation in the law," in *JURIX-2014* (Amsterdam: IOS), 81–90.
- Governatori, G., Rotolo, A., and Sartor, G. (2021). "Logic and the law: philosophical foundations, deontics, and defeasible reasoning," in *Handbook of Deontic Logic and Normative Systems*, Volume 2, eds D. Gabbay, J. Horty, and X. Parent (London: College Publications), 657–764.
- Hage, J. C. (1997). *Reasoning with Rules: An Essay on Legal Reasoning and Its Underlying Logic*. Dordrecht: Kluwer. doi: 10.1007/978-94-015-8873-7
- Hart, H. L. A. (1994). *The Concept of Law*. Oxford: Clarendon Press.
- Hart, H. L. A., and Honoré, T. (1959). *Causation in Law*. Oxford: Clarendon.
- Horty, J. F. (2011). Rules and reasons in the theory of precedent. *Legal Theory* 10, 1–33. doi: 10.1017/S1352325211000036
- Kampik, T., Gabbay, D., and Sartor, G. (2021). "The burden of persuasion in abstract argumentation," in *Clar-01* (Cham: Springer). doi: 10.1007/978-3-030-89391-0_13
- Liao, B., and van der Torre, L. (2020). "Explanation semantics for abstract argumentation," in *Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020, volume 326 of Frontiers in Artificial Intelligence and Applications*, eds H. Prakken, S. Bistarelli, F. Santini, and C. Taticchi (Amsterdam: IOS Press), 271–282.
- Liepina, R., Sartor, G., and Wyner, A. (2020). Arguing about causes in law: a semi-formal framework for causal arguments. *Artif. Intell. Law* 28, 69–89. doi: 10.1007/s10506-019-09246-z
- Lipton, P. (1990). Contrastive explanation. *R. Inst. Philos. Suppl.* 27, 247–266. doi: 10.1017/S1358246100005130
- Liu, X., Lorini, E., Rotolo, A., and Sartor, G. (2022a). "Modelling and explaining legal case-based reasoners through classifiers," in *Legal Knowledge and Information Systems - JURIX 2022: The Thirty-fifth Annual Conference, Saarbrücken, Germany, 14-16 December 2022, Volume 362 of Frontiers in Artificial Intelligence and Applications*, eds E. Francesconi, G. Borges, and C. Sorge (Amsterdam: IOS Press), 83–92. doi: 10.3233/FAIA220451
- Liu, X., Lorini, E., Rotolo, A., and Sartor, G. (2022b). "Modelling and explaining legal case-based reasoners through classifiers," in *Proc. JURIX 2022* (Amsterdam: IOS Press).
- MacCormick, D. N. (1978). *Legal Reasoning and Legal Theory*. Oxford: Clarendon.
- MacCormick, D. N., and Summers, R. S. (eds) (1991). *Interpreting Statutes: A Comparative Study*. Hanover, NH: Dartmouth.
- Maher, M. J. (2001). Propositional defeasible logic has linear complexity. *Theory Pract. Log. Program.* 1, 691–711. doi: 10.1017/S1471068401001168
- Maranhão, J., de Souza, E. G., and Sartor, G. (2021). "A dynamic model for balancing values," in *ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21–25, 2021*, eds J. Maranhão, and A. Z. Wyner (New York, NY: ACM), 89–98.
- Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007
- Miller, T., Hoffman, R., Amir, O., and Holzinger, A. (eds) (2022). *Artificial Intelligence Journal: Special Issue on Explainable Artificial Intelligence (XAI)*, Volume 307. doi: 10.1016/j.artint.2022.103705
- Peczenik, A. (1989). *On Law and Reason*. Dordrecht: Kluwer. doi: 10.1007/978-1-4020-8381-5
- Perelman, C., and Olbrechts-Tyteca, L. (1969). *The New Rhetoric: A Treatise on Argumentation*. South Bend, IN: University of Notre Dame Press.
- Pitt, J. C. (1988). *Theories of Explanation*. Oxford: Oxford University Press.
- Pollock, J. L. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge, MA: MIT. doi: 10.7551/mitpress/1887.001.0001
- Prakken, H. (2010). An abstract framework for argumentation with structured arguments. *Argum. Comput.* 1, 93–124. doi: 10.1080/19462160903564592
- Prakken, H., and Ratsma, R. (2022). A top-level model of case-based argumentation for explanation: formalisation and experiments. *Argum. Comput.* 13, 159–194. doi: 10.3233/AAC-210009
- Prakken, H., and Sartor, G. (1998). Modelling reasoning with precedents in a formal dialogue game. *Artif. Intell. Law* 6, 231–287. doi: 10.1023/A:1008278309945
- Prakken, H., and Sartor, G. (2015). Law and logic: a review from an argumentation perspective. *Artif. Intell.* 227, 214–245. doi: 10.1016/j.artint.2015.06.005
- Prakken, H., and Sartor, G. (2023). "A formal framework for combining legal reasoning methods," in *ICAIL 2023* (New York, NY: ACM).
- Prakken, H., and Vreeswijk, G. (2002). "Logics for defeasible argumentation," *Handbook of Philosophical Logic*, Volume 4, eds D. M. Gabbay, and F. Guenther (Dordrecht: Kluwer), 218–319.
- Prakken, H., and Sartor, G. (2006). "Presumptions and burdens of proof," in *Proceedings of the Nineteenth Annual Conference on Legal Knowledge and Information Systems (JURIX)*, ed T. Van Engers (Amsterdam: IOS), 176–185. doi: 10.2139/ssrn.963761
- Rissland, E. L., and Ashley, K. D. (1987). "A case-based system for trade secrets law," in *Proceedings of the First International Conference on Artificial Intelligence and Law (ICAIL)* (New York, NY: ACM), 60–66. doi: 10.1145/41735.41743
- Riveret, R., Prakken, H., Rotolo, A., and Sartor, G. (2008). "Heuristics in argumentation: a game-theoretical investigation," in *Computational Models of Argument. Proceedings of COMMA-08* (Amsterdam: IOS), 324–335.
- Roth, B., Riveret, R., Rotolo, A., and Governatori, G. (2007). "Strategic argumentation: a game theoretical investigation," in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Law* (New York, NY: ACM), 81–90. doi: 10.1145/1276318.1276333
- Rotolo, A., Governatori, G., and Sartor, G. (2015). "Deontic defeasible reasoning in legal interpretation: two options for modelling interpretive arguments," in *Proceedings of the 15th International Conference on Artificial Intelligence and Law (ICAIL'05)* (New York, NY: ACM), 99–108. doi: 10.1145/2746090.2746100
- Sartor, G. (2005). *Legal Reasoning: A Cognitive Approach to the Law*. Cham: Springer.
- Sartor, G. (2023). *Interpretation, Argumentation, and the Determinacy of Law*. San Francisco, CA: Ratio Juris. doi: 10.1111/raju.12389
- Schroeder, M. (2005). Cudworth and normative explanations. *J. Ethics Soc. Philos.* 1, 1–28. doi: 10.26556/jesp.v1i3.15
- Toni, F. (2013). A generalised framework for dispute derivations in assumption-based argumentation. *Artif. Intell.* 195, 1–43. doi: 10.1016/j.artint.2012.09.010
- Väyrynen, P. (2021). Normative explanation and justification. *Noûs* 55, 3–22. doi: 10.1111/nous.12283
- Verheij, B., Bex, F., Timmer, S., Vlek, C., Meyer, J.-J., Renooij, S., et al. (2016). Arguments, scenarios and probabilities: connections between three

normative frameworks for evidential reasoning. *Law Probab. Risk.* 15, 35–70. doi: 10.1093/lpr/mgv013

Walton, D. (2002). *Legal Argumentation and Evidence. Legal Argumentation and Evidence.* University Park, PA: Pennsylvania State University Press. doi: 10.1023/A:1021108016075

Walton, D. (2005). *Dialectical Explanation in AI.* Berlin, Heidelberg: Springer Berlin Heidelberg, 173–212.

Walton, D., Macagno, F., and Sartor, G. (2021). *Statutory Interpretation. Pragmatics and Argumentation.* Cambridge: Cambridge University Press. doi: 10.1017/9781108554572

Walton, D. N., Reed, C., and Macagno, F. (2008). *Argumentation Schemes.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511802034

Wambaugh, E. (1894). *The Case Study.* Boston, MA: Little Brown and Co.

Appendix: proof of Theorem 1

Proof. Rule-based grounded semantics are characterized by Defeasible Logic with ambiguity propagation (DL_p) (Antonioni et al., 2001), and so we know that, given any argumentation theory D and for any conclusion ψ , $D \vdash_{DL_p} \psi$ (resp. $D \not\vdash_{DL_p} \psi$) iff there exists an argument A in $GE(D)$ such that $\text{Conc}(A) = \psi$ (resp. there exists no argument A in $GE(D)$ such that $\text{Conc}(A) = \psi$) under grounded semantics (Governatori et al., 2004, Theorem 3.12). Accordingly, we can resort, with some minor modifications, to the proof developed by Governatori et al. (2022b) and which is based on the proof-theoretic properties of Defeasible Logic. We show that the complement of the considered problem is NP-complete. Namely, given the argumentation theory and the normative case, the problem is to show that the case is not stable. Hence, we have to show that a superset of the explanation that does not prove the target literal exists using the proof theory described by Governatori et al. (2004). As usual, the proof consists of two parts. Given an oracle that guesses a theory where the set of facts is a superset of the one corresponding to the explanation, we can check polynomially whether this theory proves the target literal or not [which is a standard result of Defeasible Logics (Maher, 2001)]. For the second part, we provide a polynomial encoding of 3-SAT, and we demonstrate that if the theory encoding the 3-SAT instance is not stable, then the 3-SAT instance is satisfiable. A 3-SAT instance is given by

$$\bigwedge_{i=1}^n \phi_i$$

where $\phi_i = \psi_i^1 \vee \psi_i^2 \vee \psi_i^3$. Its encoding in Defeasible Logic is given by the argumentation theory $D = (R, \emptyset, \emptyset)$ where R contains, for every clause ϕ_i , the following rules⁶:

$$r_{ij}: \psi_i^j \Rightarrow \phi_i \quad j \in \{1, 2, 3\}$$

plus the two rules:

$$\begin{aligned} r_{sat}: \phi_1, \dots, \phi_n &\Rightarrow sat \\ r_{nsat}: &\Rightarrow \neg sat \end{aligned}$$

The encoding is polynomial in the size of the 3-SAT instance. We consider the case given by the empty set of facts and $\neg sat$ as the target literal. It is immediate to verify that $D \vdash_{DL_p} \neg sat$: r_{nsat} is the only applicable rule. The set of admissible facts (see Definition 3) consists of all literals ψ_i^j and $\neg \psi_i^j$. To show that \emptyset is not stable we have to find a subset of admissible facts C such that $D' = (R, C, \emptyset) \not\vdash_{DL_p} \neg sat$.⁷ For a (consistent) set of admissible facts C , we build the interpretation I as follows:

$$I(\psi_i^j) = \begin{cases} TRUE & \psi_i^j \in C \\ FALSE & \text{otherwise} \end{cases}$$

We cannot show that $D' \not\vdash_{DL_p} \neg sat$ iff $I \models \bigwedge_{i=1}^n \phi_i$. To disprove $\neg sat$, the rule r_{sat} has to be applicable. This means we need to prove ϕ_i . This implies that for each ϕ_i at least one of the rules $r_{i,1}$, $r_{i,2}$ and $r_{i,3}$ is applicable too. Consequently, one of ψ_i^1 , ψ_i^2 , and ψ_i^3 is derivable. Given there are no rules for ψ_i^j , ψ_i^j is provable iff $\psi_i^j \in C$. Accordingly, $I(\psi_i^j) = TRUE$. Thus, for every clause we have an element in it that makes the clause true, thus $I(\phi_i) = TRUE$, for every i and so the 3-SAT instance is satisfiable. Conversely, when $I \models \bigwedge_{i=1}^n \phi_i$, $I \models \phi_i$ for every $1 \leq i \leq n$. Thus, for each ϕ_i , there is a ψ_i^j such that $I(\psi_i^j) = TRUE$, and so $\psi_i^j \in C$. Therefore, $D' \vdash_{DL_p} \psi_i^j$, from which we derive that for every i , $D' \vdash_{DL_p} \phi_i$, making r_{sat} applicable, which implies $D' \not\vdash_{DL_p} \neg sat$.

Of course, the following holds as well.

Theorem 2. Given a theory D and an explanation by sufficient facts \mathcal{F} relative to D , the problem of determining if \mathcal{F} is not stable is NP-complete.

⁶ Notice that we use ϕ_i as a variable for a clause in the 3-SAT instance and as a literal (representing the clause) in the corresponding defeasible logic encoding.

⁷ More precisely, we have to constructively disprove such a conclusion (i.e., we have to constructively show that there is no proof), something that Defeasible Logic support in its proof theory.



OPEN ACCESS

EDITED BY
Loizos Michael,
Open University of Cyprus, Cyprus

REVIEWED BY
Ute Schmid,
University of Bamberg, Germany
Tarek R. Besold,
Eindhoven University of Technology,
Netherlands

*CORRESPONDENCE
Selmer Bringsjord
✉ selmerbringsjord@gmail.com

RECEIVED 14 January 2023
ACCEPTED 04 October 2023
PUBLISHED 08 January 2024

CITATION
Bringsjord S, Giancola M, Govindarajulu NS,
Slowik J, Oswald J, Bello P and Clark M (2024)
Argument-based inductive logics, with
coverage of compromised perception.
Front. Artif. Intell. 6:1144569.
doi: 10.3389/frai.2023.1144569

COPYRIGHT
© 2024 Bringsjord, Giancola, Govindarajulu,
Slowik, Oswald, Bello and Clark. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Argument-based inductive logics, with coverage of compromised perception

Selmer Bringsjord^{1*}, Michael Giancola¹,
Naveen Sundar Govindarajulu¹, John Slowik¹, James Oswald¹,
Paul Bello² and Micah Clark³

¹Rensselaer AI & Reasoning (RAIR) Lab, Department of Computer Science, Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY, United States, ²Naval Research Laboratory, Washington, DC, United States, ³College of Information Sciences and Technology, Pennsylvania State University, State College, PA, United States

Formal deductive logic, used to express and reason over declarative, axiomatizable content, captures, we now know, essentially all of what is known in mathematics and physics, and captures as well the details of the proofs by which such knowledge has been secured. This is certainly impressive, but deductive logic alone cannot enable rational adjudication of arguments that are at variance (however much additional information is added). After affirming a fundamental directive, according to which argumentation should be the basis for human-centric AI, we introduce and employ both a deductive and—crucially—an inductive *cognitive calculus*. The former cognitive calculus, *DC&C*, is the deductive one and is used with our automated deductive reasoner ShadowProver; the latter, *IDC&C*, is inductive, is used with the automated inductive reasoner ShadowAdjudicator, and is based on human-used concepts of *likelihood* (and in some dialects of *IDC&C*, probability). We explain that ShadowAdjudicator centers around the concept of competing and nuanced arguments adjudicated non-monotonically through time. We make things clearer and more concrete by way of three case studies, in which our two automated reasoners are employed. Case Study 1 involves the famous Monty Hall Problem. Case Study 2 makes vivid the efficacy of our calculi and automated reasoners in simulations that involve a cognitive robot (PERI.2). In Case Study 3, as we explain, the simulation employs the cognitive architecture ARCADIA, which is designed to computationally model human-level cognition in ways that take perception and attention seriously. We also discuss a type of argument rarely analyzed in logic-based AI; arguments intended to persuade by leveraging human deficiencies. We end by sharing thoughts about the future of research and associated engineering of the type that we have displayed.

KEYWORDS

inductive logic, compromised perception, argument and automated reasoning, Monty Hall dilemma, cognitive robotics, AI

1 Introduction

Formal deductive logic, used to express and reason over declarative, axiomatizable content, captures, we now know, essentially all of what is known in mathematics and physics, and captures as well the details of the proofs by which such knowledge has been secured. This is impressive certainly, but even simple scenarios explain a very different story: for example, if (human) Alice perceives a blue cube on a table, then accordingly declares that she believes that there is a blue cube thereon, while Bob, beside her and looking also at the table through his pair of glasses, asserts “No, actually that’s an orange sphere,” deductive logic alone

cannot enable rational adjudication of the disagreements between them. The great pioneer of modern inductive logic, Rudolph Carnap, fully understood this in the mid-20th century during the heyday period of deductive logic brought about principally by Gödel. Carnap would say, and the logicians and mathematicians today who continue his vibrant legacy in the form of what is known as *pure inductive logic* (PIL) (Paris and Vencovská, 2015) would still say, that “There is a blue cube on the table” and “There is an orange sphere on the table” should each be assigned a probability value (a real number between 0 and 1, inclusive), and this content, combined with additional probabilized propositions, can be used in a process that dictates what should be rationally believed. Unfortunately, Carnap and his followers pay little attention to the “coin of the realm” in human reasoning and decision-making: arguments and argumentation.¹ This tradition (which began before Carnap and includes e.g., Keynes, 1921) also runs afoul of the brute fact that humans very rarely use probabilities and the probability calculus (and when they are “boxed in” to using probabilities, very rarely use them correctly, as shown by the infamous “Linda Problem”, nicely discussed in this connection by Kahneman, 2013). In addition, this tradition in inductive logic avoids the brute fact that Alice and Bob, humans in general, and also today’s cognitive robots, inevitably perceive in messy environments that render percepts highly uncertain (e.g., what are the ambient lighting conditions in the room Alice and Bob are in?). We introduce below a family of novel inductive logics, based centrally on human-used concepts of *likelihood*, that center around the concept of competing, nuanced arguments adjudicated through time. We present three case studies in which likelihood is key: Case Study 1 involves the famous Monty Hall Problem.² Case Study 2 makes vivid the efficacy of our calculi and automated reasoners in simulations that involve the robot (PERI.2). In Case Study 3, as we explain, the simulation employs automated reasoners joined with the cognitive architecture ARCADIA, which is designed to computationally model cognition in ways that take perception and attention seriously. Penultimately, we discuss a class of arguments hitherto largely ignored in logicist AI, such as arguments designed to persuade despite the fact that they are unsound. We end by sharing thoughts about the future of research and associated engineering of the type that we have displayed herein.

The remainder of the present study unfolds as follows. In the next Section 2, we explain, affirm, and (albeit briefly) defend

our “prime directive,” in a word that argumentation must be the basis of human-level, and human-centric, AI. Next, we (Section 3) briefly point out that, putting it mildly, perception has not exactly been treated in a deep way in the history of logicist AI—despite the fact that immediately instructive parables such as the Alice-Bob sketched above have been obvious since McKeon (1941)³ presented to humanity, in his *Organon*, the first formal logic, with algorithms for determining whether arguments expressed therein are formally valid.⁴ What follows is a section devoted to giving an historical perspective on our research (Section 4) and coverage of a considerable amount of related prior study. The next section lists the specific desiderata for argument-centric automated defeasible (= non-monotonic) reasoning that we seek and abide by and which are satisfied by the logico-mathematics, systems, and case-study demonstration we present herein. We then (Section 6) orient the reader to our brand of logicist AI by briefly explaining our background logico-mathematics; this section ends with a subsection in which the specifications for the two pivotal cognitive calculi alluded to above (*DC&C* & *IDC&C*) are given. Section 7 presents, in turn, the three case studies we have promised above. The penultimate section of the study is devoted to explaining a category of arguments premeditatedly designed to be unsound but (in fact in some cases more) persuasive. In our final Section 9, we touch upon the need to solve paradoxes in the intersection of reasoning and perception, point out that future study is needed to address pictorial arguments (which are common in the human case), and offer a few final remarks.

2 Argumentation must ground human-centric AI

We believe that the basis for rational human use of AI technology is, or at least ought to be, argumentation, computationally treated, and managed. In this regard, we wholly concur with Dietz et al. (2022). For us, this is a firm and fundamental directive that guides our research. For convenient reference to this directive in the remainder of the present study, we refer to it as simply ‘Dir’. Notably, we stipulate that Dir specifies for us *rational* human use of AI. Obviously, there are irrational uses of AI that, by definition, make argumentation decidedly unwanted, for at least some of the humans involved. For instance, Jones may wish to simply make, activate, and then violently destroy AI technology (because he is in the grip of an pathological level of hatred of all things both artificial and human-like), and it is exceedingly hard to observe how this non-cerebral use of AI should be mediated by argumentation.⁵ Of course, we anticipate that most human use of AI technology will indeed be rational.

1 As well as proofs, which we take to be just a special case of arguments. Abstractly put, an argument for us is a sequence of formulae in some formal language of some logic or logics, where the sequence links these formulae by instances of inference schemata. A proof is an argument in which (i) the inference schemata in play are restricted to deductive ones, and (ii) some premises given in the sequence in question enjoy special status because they are members of a pre-identified axiom system (e.g., axioms for Euclidean geometry, or for arithmetic, or topology).

2 MHP, as a matter of fact, in our formalization and solution, involves *both* likelihood and probability. Since the emphasis, herein, is very much on the former, we do not bring to bear our full formalization and implementation of the probability calculus of Kolmogorov (1933) within a richer version of *IDC&C*. Doing so would be overkill in the present study, since the key manner of handling uncertainty is here cognitive likelihood not probability.

3 A modern translation into English of Aristotle’s writings.

4 A nice, modern overview of this fragment of first-order logic = \mathcal{L} is given in the study mentioned in Smith (2017).

5 The Spielberg–Kubrick film *A.I.* includes a rather depressing depiction of a number of humans who are, in fact, like Jones. We refer to the (disturbing) stretch of the film in which humans destroy robot after robot in violent, sadistic fashion.

So far, we have referred to “AI technology.” Let us be a bit more accurate, by speaking of **artificial agents**, in accordance with the comprehensive, respected textbooks for the field of AI (see Luger, 2008; Russell and Norvig, 2020). In these studies, in broad strokes, which suffice for the present study, artificial agents, located in a given environment, take in percepts of that environment as input and compute in some fashion over this input (along with various information from other sources and of other types), and this computation leads them to perform actions as output. In our approach, that of logic-based/logicist AI, the computation that maps percepts to actions is specifically that of automated reasoning, and the performance of all actions is the result of a conclusion reached by inferences which are, in each and every case, formally verified (which means that in the case of actions carried out by our logicist artificial agents in the coming trio of case studies, correctness is invariably proved).⁶

Next, and importantly, we point out that **Dir** is not just randomly pulled from thin air: we follow it because not doing so at best makes rational human use of artificial agents less productive and at worst makes such use in some cases outright dangerous. This holds true not only when the artificial agents in question operate in a manner divorced from the type of AI that intimately connects to argumentation (i.e., logic-based AI, to which we are adherents and which grounds the new research we present below) but also when these agents are in fact logic-based (or logicist). We explain this now with an example of each of these two types of cases.

2.1 The need for argumentation in non-logicist systems for rational human use

To observe the desirable role of argumentation in an example of dangerous human use of artificial agents engineered in the absence of logicist formalisms and techniques, we can consider the logic-less “large language model” Galactica, engineered and released by its creators in order to provide human beings with “a new interface for science” (Taylor et al., 2021), at least seemingly a rather laudable goal for human-centric AI.⁷ As a matter of fact, Galactica, with minimal prompts from a human, can quickly write entire scientific papers, replete with references. It does this by way of deep learning only. Unfortunately, when used by some human scientists, Galactica simply concocted many things having no relation to relevant reality. For instance, some of the references in scientific papers it “wrote” were completely fictional but of course sounded quite legitimate. The debacle, efficiently chronicled in the study mentioned in Heaven (2022), shows that Galactica poses the danger of unethical submission of scientific papers that appear sound yet are anything but. In short, a Galactic-written paper is—to use the adjective we flesh out in the study mentioned in Section 8—sophistic.

⁶ This not being a study on formal verification, we omit formal verification.

⁷ Exactly parallel points as we make in the present section could be made about GPT-4, discussed (and greatly lauded) in the study mentioned in Bubeck et al. (2023). The details behind GPT-4 are unfortunately proprietary; Meta has made available to all its transformer infrastructure.

What is the solution? The solution is that the actions taken by artificial agents, in this case the assembling of scientific papers on the basis of purely statistical processing of historical data, be intimately tied to checkable arguments in support of what is expressed in said papers. As we explain below, in our argument-based AI, all outputs are the result of automatically found proofs and/or formal arguments; and these proofs and arguments can not only be inspected by humans but can be *certified* by artificial agents that automatically check these proofs/arguments.

2.2 The need for argumentation in logicist systems for rational human use

What about artificial agents in the second kind of case? That is, what about artificial agents that are in fact logic-based, but argumentation does not mediate between the humans using such agents and the agents’ actions? An illuminating example to consider here is the famous “Monty Hall Problem” (MHP), which is going to be a bit of a theme in the present study, and which, following the study mentioned in Bringsjord et al. (2022b), we sum up as follows:

The (3-door) Monty Hall Problem (MHP₃)

Jones has come to a game show and finds himself thereon selected to play a game on national TV with the show’s suave host, Monty Hall. Jones is told correctly by Monty that hidden behind one of three closed, opaque doors facing the two of them is \$1,000,000 USD, while behind each of the other two is a not-exactly-clean, obstreperous donkey whose value on the open market is charitably pegged at \$1. Monty reminds Jones that this is a game and a fair one, and that if Jones ends up selecting the door with \$1M behind it, all that money will indeed be his. (We can assume without loss of generality that Jones’ net worth has nearly been exhausted by his expenditures in traveling to the show.) Monty also reminds Jones that he (= Monty) knows what is behind each door, fixed in place until the game ends.

Monty asks Jones to select which door he wants the contents of. Jones says, “Door 1.” Monty then says: “Hm. Okay. Part of this game is my revealing at this point what’s behind one of the doors you didn’t choose. So . . . let me show you what’s behind Door 3.” Door 3 opens to reveal a cute but very — economically speaking — unsavory donkey. Monty now to Jones: “Do you want to switch to Door 2, or stay with Door 1? You’ll get what’s behind the door of your choice, and our game will end.” Monty looks briefly into the camera, directly.

What should Jones do if he is logical?

Unfortunately, as nicely explained in the study mentioned in Friedman (1998) and many other papers and books, including the recently published *Rationality* from Pinker (2021), the vast majority of humans respond by saying that Jones should not switch. In fact, as the history of MHP₃ has shown, many mathematicians aggressively insisted that the rational policy is STAY, not SWITCH.⁸

⁸ See Tierney (1991), and for a shorter historical account, Chap. 1 of Pinker (2021). For fuller discussion of the mathematicians to whom we have just referred, see Granberg (2014).

However, the provably correct response to the question is that Jones should follow a SWITCH policy.

Now, suppose that some artificial agents have perceived the MHP₃ problem, automatically discovered the correct answer, and now share that answer with a typical human who fails to grasp the problem and thought that the correct policy was STAY. How helpful is this artificial agent going to be to this human? Not very. After all, the human does not know *why* the correct answer is SWITCH. The obvious solution, given the need for genuinely helpful human-centric AI, is a class of artificial agents that can not only find solutions but also provide cogent, compelling, verified arguments certifying those solutions. If such a thing is provided in the present case, the human will be enlightened. As will be observed later in the study, this is what our artificial agents can do, even in cases where the percepts to these agents are “clouded.”

3 The perception lacuna/challenge

The lead author has been carrying out logicist AI R&D for three decades and can count, on one hand, systems that genuinely integrate automated reasoning with the full gamut of the main human-level cognitive operators, and with attention and perception understood in keeping with state-of-the-art cognitive science. It is even harder to find such systems that are rigorous and computationally implemented. This missing type of research is what the present section’s heading refers to as a lacuna.

Addressing this inadequacy is observed as important by others. For example, Dietz et al. (2022), when setting out desiderata for HCAI systems, include that such systems must exhibit a “body-mind like model of operation to sense, recognize, think, and act” (Dietz et al., 2022). For us, broadly speaking, here, “think” is constituted by reasoning,⁹ and we associate “sense, recognize” with attention and perception. Later, in the same study, when discussing what is needed for true success in HCAI and indeed for any brand of AI overall that aspires to cover the human-level case, Dietz et al. (2022) point to the following challenge:

[Such success must include showing] how the internal integrated operation of cognition, from low-level perception to increasingly higher levels of cognition, is supported by an appropriate architecture, and how an individual’s cognition is integrated with the external physical and social environment (Dietz et al., 2022; emphasis ours).

In keeping with such demands, we are actively working on the integration of attention and perception with (esp. rational) human-level reasoning, in a manner that takes account of a given artificial agent’s external and physical environment.¹⁰ Another way

9 A view affirmed and used in the Prolog-centric (and hence only extensional) and pedagogically oriented tour through thinking as reasoning in the study mentioned in Levesque (2012).

10 The lead author, along with author P. Bello, is, in addition, convinced not only that an agent’s perception of its internal, mental environment is equally important, but also working on formalizing and implementing the relationship between internal perception (which calls for its own intensional operator in cognitive calculi) and reasoning. For an exploration of internal

to put our goal of integration is to say that it is aimed at unifying “bands of cognition.” This aim is characterized by the following instructive quote:

Interestingly, [the] missing convergence toward unified theories of cognition persists across and within the bands of cognition Newell (1990). Bridging the gap between Newell’s bands of cognition still exists as a problem and the main challenge remains. How do we organize the internal processes of a system at different levels such that they can operate internally linking perception and high-level cognition, by facilitating their meaningful integration with other systems and the external human participating environment? (Dietz et al., 2022).

The question in the last sentence of this quote is fundamentally what drives our integration of our automated-reasoning systems with perception; and below, Case Studies 2 (Section 7.2) and 3 (Section 7.3) demonstrate some of this study.

We now turn specifically to the latest version¹¹ of our desiderata for human-level argumentation (and proof) systems, specified and implemented within the constraints of our particular approach to human-level logicist-AI.

4 Historical context and related work

In the present section, we first provide some historical contexts (Section 4.1) and then (Section 4.2) summarize related studies to set the stage for giving our own specific desiderata, which drive our work.

4.1 Historical context

Sensible presentation of our desiderata for an argument-centric automated reasoner must, at least to some degree, be contextualized historically. We, thus, now issue some remarks along this line. Needless to say, these remarks will not constitute a full history of systematic, precise work in argumentation-based formal and/or computational logic.¹²

perception in self-conscious robots that is integrated with robust reasoning in a cognitive calculus, see Bringsjord et al. (2015).

11 Ancestors and less-developed lists of the desiderata in question have been given in some previous studies, including e.g., Bringsjord et al. (2020a).

12 A comprehensive history, in our opinion, needs to be composed by someone at some point. Part of the challenge is the need for the brave author who takes this project to be fluent in at least both ancient Greek and German. The former language is key because Aristotle can be viewed as the primogenitor (e.g., see *Topica* and *De Sophisticis Elenchis* in his *Organon*, available in McKeon, 1941). German is crucial because, even to this day, the remarkable work of Lorenzen, momentarily discussed, has not been fully translated from the German. In addition, the lead author is personally of the view that the work of Leibniz in formal logic (which includes the long-before-Frege invention of both first-order logic and modal logic; see Lenzen, 2004), and in particular work toward his dream of a “universal logic” (which is expressly intended to cover the dynamic argumentation of multiple,

From an historical perspective, our approach, formalisms, and AI technology for argumentation can be viewed as having roots in *dialogue logic*, the seminal 1958 introduction of which, in formal terms, is due to Lorenzen (1960). As Walton and Krabbe (1995) have pointed out, Lorenzen's work can be traced to treatment of dialogue in Aristotle (and in this connection see note 12). Since an excellent and remarkably efficient summary of dialogue logic/games is provided by Bench-Capon and Dunne (2007), a paper to which we shall return to, and rely upon, later, there is really little it makes sense for us to recapitulate from the dialogue tradition. We make only three quick points, as follows:

1. When one considers a two-person dialogue game in which Proponent and Opponent struggle over some proposition, our ShadowAdjudicator can be viewed as the judge charged with rendering rulings as to the winner.
2. We allow any number of agents to articulate and propose arguments on the proposition at hand (a fact that becomes concretized in our case studies).
3. Our third point is by far the most important of the present trio and consists of our explicitly affirming an insight into Bench-Capon and Dunne (2007), which in a word is that the specification of the internal structure of arguments, vis-à-vis conformity to inference schemata,¹³ is crucial. This insight is, in fact, explicitly included as a desideratum in **Des**, as shall be soon observed. In our case, inference schemata, as will be clearly and concretely observed in the cognitive calculi we bring to bear in our case studies, are not only formal (as is the case even with something is straightforward as *modus ponens*) but also are intensional in nature and parameter-rich (e.g., the inference schemata specified for both *DCEC* and *IDCEC* given in Section 6.2.1).¹⁴

interacting agents), is quite relevant to any full history of the sort we are imagining, which means that command of Latin and French also becomes a requirement for the relevant scholarship [we return to the Leibnizian nature of our paradigm below (Section 4.2)].

13 We read as follows:

It has been seen that Dung's fundamental model, as described in [73], abstracts away such internal structure from individual argument in order to focus on the manner in which arguments interact *via* the defined attack relationship. In unfolding the exact nature of "the argument x attacks the argument y ," however, the *reason* why such an attack is present needs to be considered in terms of those *structural schema* underlying the arguments x and y from which the attack arises. Such an interpretation, therefore, raises issues that concern the form an argument might take, i.e. issues regarding the components and representation of *arguments* rather than the process and outcome of the *argumentation* involved (Bench-Capon and Dunne, 2007, p. 625).

14 It is worth pointing out that Dung's (1995) abstract scheme for arguments is (unbeknown in our experience to most researchers working in AI and computational argumentation systems and formalisms) related to, indeed in some non-trivial respects anticipated by, a longstanding sub-area of formal logic that spans both extensional and intensional logic; we refer to *justification logic*. A nice starting point is Artemov (2008). The core idea in justification logic (to simplify rather harshly for economy) is that formulae

Turning now to more "classical" logicist work in 20th-century AI, we begin by rehearsing that, as the reader will likely recall, standard first-order logic \mathcal{L}_1 (and all its fragments, such as the propositional calculus and zero-order logic \mathcal{L}_0 ¹⁵) is *monotonic*: the arrival of new information cannot change the result of previous inferences. That is,

$$\text{If } \Phi \vdash \phi \text{ then } \Phi \cup \Psi \vdash \phi,$$

where Φ, Ψ are sets of formulae in the formal language of \mathcal{L}_1 , and ϕ is an individual formulae in this logic; we implicitly universally quantify over these three elements. In stark contrast, defeasible reasoning is *non-monotonic*. It has long been known in AI that such reasoning is desirable when formalizing much real-world reasoning. For instance, there are the early, classic default logics of Reiter (1980), in which epistemic possibilities hold in default of information to the contrary. In general, it is desirable to be able to reason based on beliefs which could potentially be false, and to be able to retract such beliefs when new, countervailing information arrives. Our coming desiderata **Des** call for more than this. Default logic, despite having many virtues, does not satisfy **Des**; the reason, in short, is that it has no provision for intensional/modal operators corresponding to cognitive verbs known to stand at the heart of human-level cognition (such as *believes*, *knows*, *perceives*, and *communicates*), as cognitive psychologists have shown (for an overview, see Ashcraft and Radvansky, 2013). These verbs are also known as *propositional attitudes* by logicians and philosophers, and their inclusion in a given logic makes that logic an *intensional* one, not just an *extensional* one (Fitting, 2015; Nelson, 2015).

A diagnosis parallel to that issued for default logic holds with respect to circumscription, an impressive non-monotonic form of reasoning introduced long ago by McCarthy (1980). Circumscription makes no provision for modal operators to

that express some proposition, say ϕ , are accompanied by justifications, and we thus have for instance $t : \phi$, where t is the justification. Justifications, here, have long been conceived as proofs and/or arguments. This tradition, and the logico-mathematical results that have been obtained, run back to a time (circa 1930 in the case of extensional logic, within mathematical logic) quite before the study by Dung and others on abstract schemes for the systematic study of argumentation. For a detailed overview, see Artemov and Fitting (2020) (while this overview credits some early mathematical logicians, e.g., Kolmogorov, 1932, with laying the foundations of justification logic because of their identifying "truth" with "provability," it does not credit, as the first author of the present study feels it should, those who established proof-theoretic semantics, as also laying part of these foundations. As observed below when we present the technical basis of our approach to and work on computational inductive logic, proof-theoretic semantics has greatly influenced this approach/work). Regarding our own approach, the lack of internal structure in justifications in justification logic, which parallels the situation with respect to Dung's approach, means that our computational logics and AI systems for argumentation-centric AI are very different. This is expressed explicitly in desiderata d_4 and d_5 in our set **Des** of desiderata, given in Section 5.

15 No quantifiers, but constants to denote individuals, unrestricted use of n -ary relation and function symbols, the identity = relation, and inference schemata for deduction using identity, e.g., that if $a=b$ and $\phi(a)$ (a formula in which constant a occurs), then inference to $\phi(b)$ is permitted.

capture cognitive attitudes and does not include the type of human-digestible arguments we require. There have been defeasible-reasoning models and systems that do include arguments that compete against each other in a manner at least somewhat similar to our concept of adjudication. The closest case in point is the work of Pollock (1995). For an excellent survey of defeasible reasoning systems that are, at least to some degree, argument-based (see Prakken and Vreeswijk, 2001).¹⁶

4.2 Related work

Argumentation in AI, as our foregoing coverage in the present section clearly indicates, is long established. To now further set the stage for enumeration of the desiderata that govern our own work, we turn to the 21st century. A truly excellent overview of this more recent work is provided by Bench-Capon and Dunne (2007), a study we have already relied upon, and which at its outset attempts to distinguish between mathematical reasoning and proofs vs. reasoning observed in arguments. The distinction is given, in part, to provide a basis for a number of studies in a special issue of *Artificial Intelligence* that follow this study, and as far as we can determine from reading these other studies, the distinction is affirmed by all the authors. However, while we certainly acknowledge that this foundational distinction is widely affirmed, it is not one that applies to our approach. In a word, the reason is that inductive logic, computationally treated, as has been explained by the lead author elsewhere (see Bringsjord et al., 2021, 2023b), must conform to the Leibnizian dream of a “universal logic” that would serve to place rigorous argumentation (in e.g., even jurisprudence) in the same machine-verifiable category as mathematical reasoning. This means that the fundamental distinction made in the study mentioned in Bench-Capon and Dunne (2007), while nearly universally accepted, does not apply to the approach taken herein. In particular, our desideratum d_5 given in the next section treats proof and argument the same in this regard: both are formally, mechanically verifiable. We now take a closer look at these matters.¹⁷

16 For an efficient overview of defeasible reasoning, in general, the interested reader for whom defeasible/non-monotonic reasoning is new is directed first to the study mentioned in Koons (2017).

17 There are at least two other important, substantive matters that must for economy be left aside, which are quite important. The first is that as a matter of fact, the arguments and proofs that are key to our study are often expressed in what is as far as we know a novel graphical form of so-called “natural reasoning”: *hypergraphical* natural reasoning because arguments, proofs, and semantic structures [e.g., a hypergraphical version of so-called “truth trees” (as nicely introduced in Bergmann et al., 2013)] are all expressed as hypergraphs (Berge, 1989; Bretto, 2013), including 3-dimensional hypergraphs; see Bringsjord et al. (2023a). We observed our hypergraphical approach as being within the general fold of graphical schemes for argumentation, a nice example of which is given in the study mentioned in Bench-Capon et al. (1992), which is, in turn, within the general approach of Toulmin (2003). A look at a robust hypergraphical proof within a logicized theory of special relativity that faster-than-light travel is impossible (see Govindarajulu et al., 2015).

Bench-Capon and Dunne (2007) present four properties that mathematical reasoning is said to have, but which argumentation is said to lack. We do not think that any of these properties hold of mathematical reasoning but not of argumentation; however, unsurprisingly, full analysis is beyond the present scope. We thus comment on only their fourth property, which relates directly to the issue we have already raised. This fourth property is expressed verbatim by these two authors as follows:

[I]n mathematical reasoning ... [r]easoning and conclusions are entirely *objective*, not susceptible to *rational* dispute on the basis of subjective views and prejudices. Proof is demonstration whereas argument is persuasion (Bench-Capon and Dunne, 2007, p. 620).

Our reaction is rooted in Leibniz, whose objective was explicitly to do away with mere persuasion (on weighty matters), and rational disputes were to be crisply adjudicated by computation over arguments—computation we formalize and implement as automated adjudication in our sense, displayed in the present study.¹⁸ To meet this objective, two things were needed, a universal formal/logical language, the *characteristic universalis*, and automated-reasoning technology, the *calculus ratiocinator* (Paleo, 2016). The idea is that when these are obtained, rigorous argumentation (arising from disagreements that drive the production of competing arguments) can be computationally adjudicated, and arguments can also be discovered by computation. It is not important here at all as to whether Bringsjord and Govindarajulu have in fact found, as they claim, these two things (e.g., claimed and justified by an argument, in Bringsjord et al., 2023b); the important point is that the paradigm advanced by the research and engineering, reported herein, is based on a premeditated conflation of argument/argumentation and proof/mathematical reasoning.¹⁹

A second wide-ranging treatment of reasoning in AI approached via logic is provided in the study mentioned in

18 Bench-Capon and Dunne (2007) astutely concede in footnote #2 on page 620 that even in mathematics there are disputes about premises (or axioms); they give the Axiom of Choice as an example. However, they insist that a *sine qua non* for rational dispute is having on hand an “alternative theory” (in this case e.g., ZF rather than ZFC). Moreover, at any given point in mathematics (and, needless to say, mathematical physics pursued through formal logic), there has been dispute in the absence of an alternative theory. A case in point is the rejection to Cantor’s seminal introduction of transfinite numbers and their logic by many mathematicians, on the grounds not of an alternative theory, but of their perceived near absurdity. A more recent case in point is that Gödel’s now-fulfilled prophecy that new axioms governing very large sets (e.g., the independence of the Continuum Hypothesis from ZF/ZFC) would simply be legislated. Another example, perhaps the sharpest one, is the rejection of infinitesimals in the absence of alternative theory that accommodated them (rather than simply leaving aside, as in the case of limits), and then the subsequent vindication of Leibniz on infinitesimals by Robinson (1996).

19 Alert readers will perceive that our terminological practice in the present study reflects this, as e.g., we sometimes use “proof” instead of “argument” to refer to a chain of inferences found automatically by our ShadowProver system.

Davis (2017), and we now offer analysis of it in relation to our own approach as well. Davis (2017) provides a survey of the attempt to formalize commonsense reasoning in a logic, and certainly some (including a reviewer of an earlier draft of the present study who encouraged us to factor in Davis's study) regard our argumentation-focused work in human-centric AI to be at least in large measure devoted to commonsense reasoning. It seems reasonable, for example, to view MHP as a commonsense-reasoning challenge. At any rate, for the sake of argument, we are more than willing to agree that this is the case. However, while the survey in question is as far as it goes in our opinion masterful, our approach is quite different in important, enlightening ways, as we now explain. We list three ways our work in computational inductive logics for formalization and automation of argumentation differs from all the work that Davis (2017) surveys:

1. *Our foundation is decidedly not mathematical logic.* Repeatedly, Davis writes that the approach he is analyzing and summarizing is the use of “mathematical logic” for formalizing commonsense reasoning. For example, on p. 651 he writes: “One of the most studied approaches toward [the] goal [of formalizing commonsense reasoning] has been to use formal *mathematical logic*” (emphasis ours). On p. 656 he writes: “This paper focuses on developing representations of fundamental commonsense domain by hand by experts using *mathematical logic* as a framework” (emphasis ours). There are other such quotes available in the study, but we omit them as redundant. The point, here, is that mathematical logic is the branch of logic devoted to formalizing mathematical reasoning, a pursuit that started with Aristotle (Glymour, 1992). However, our roots are in the tradition of devising formal logics that can capture human-level cognition, not mathematical reasoning or anything of the sort (see Bringsjord et al., 2023c). In a word, mathematical logic has for over two millennia been purely *extensional*.
2. *We straddle formal deductive logic and formal inductive logic; the latter is not on Davis's radar screen.* The phrase “inductive logic” (nor any equivalent) does not occur in Davis (2017). Given that the work surveyed therein is avowedly aligned with mathematical logic (as we have pointed out), this is unsurprising. However, formal logic is a large discipline that—as we have shared above—includes not just deductive logic but inductive logic, and the latter is itself any enormous enterprise now. There is, for example, no mention of the Carnapian edifice of pure inductive logic (Paris and Vencovská, 2015) in the survey, and no mention of inductive logic as the part of logic that includes analogical and abductive reasoning and enumerative induction (Johnson, 2016). To his great credit, Davis does consider logics in the categories of *non-monotonic*, *probabilistic*, and *fuzzy* (see final paragraph of p. 664). Moreover, here, there is for sure a connection to our approach and formalisms, but one important difference is that our study makes crucial use of the concept of *likelihood*, as distinct from probability (see below).
3. *There is an expressivity canyon between what Davis is concerned with vs. our cognitive calculi (= our logics).* Our cognitive calculi start at the level of quantified multi-modal logic and expand from there. However, when Davis reports on modal logics, his orientation is that of containment. For instance, he reports

with approval that “propositional modal logics . . . are often both expressive enough for the purpose at hand and reasonably tractable, or at least decidable” (p. 662). However, from the standpoint of human-level cognition, our position is that modal operators are almost invariably accompanied by quantification (and in fact quite naturally to \mathcal{L}_3).

Now, what about work specifically in defeasible argumentation systems, with an eye to the desiderata **Des** to be laid down momentarily in the next section? We wrap up the present section by summarizing two examples of such related prior study, and distinguish them from our approach in broad strokes:

1. Modgil and Prakken (2014) have presented and made available a general, computational framework—ASPIC⁺—for structured argumentation. This impressive framework is based on two fundamental principles, the second of which is that “arguments are built with two types of inference rules: strict, or deductive rules, whose premises guarantee their conclusion, and defeasible rules, whose premises only create a presumption in favor of their conclusion” (p. 31 of Modgil and Prakken, 2014). This second principle is directly at odds with desideratum d_5 in the full list **Des** given in the next section. In our approach, all non-deductive inference schemata are checked, in exactly the way that deductive inference schemata are. For instance, if some inferences are analogical in nature, as long as the schema $\frac{\Phi}{C}$ (Φ for a collection of premises in some formal language and C for the conclusion) for an analogical inference is correctly followed, the inference is watertight, not different than even *modus ponens*, where of course specifically we have $\frac{\phi \rightarrow \psi, \phi}{\psi}$.²⁰
2. Cerutti et al. (2017) is an overview of implementation of formal-argumentation systems. However, the overview is highly constrained by two attributes. The first is that their emphasis is on Turing-decidable reasoning problems, whereas our emphasis—as reflected in **Des** and in our case studies—is on reasoning challenges that, in the general case, are Turing-undecidable. As to the second attribute, the authors are careful to say that their study is constrained by the “basic requirement” that “conflicts” between arguments are “solved by selecting subsets of arguments,” where “none of the selected arguments attack each other.” Both of these attributes are rejected in our approach; in fact, in the coming trio of case studies (Section 7), automated processing is possible *because* of this rejection. With respect to the first of their attributes, most of the interesting parts of automated-reasoning science and technology for us only start with problems at the level of the *Entscheidungsproblem*; see in this regard desideratum d_7 . As to the second attribute, it is not true for our approach.

Now, as promised, here are our desiderata, which the reader will notice are in play when we reach our case studies.

²⁰ For a discussion of this sort of explicit rigidity in the case of analogical inference, see Bringsjord and Licato (2015). Analogical inference schemata arise again below, in Section 8.

5 Desiderata driving our approach

We denote the 7-fold desiderata for the capability we seek in our automatic argumentation systems by ‘Des’. An automated reasoner of the kind we seek must:

Desiderata “Des”

- d_1 be defeasible (and hence non-monotonic) in nature (when new information comes to light, past reasoning is retracted in favor of new reasoning with new conclusions);
- d_2 be able to resolve inconsistencies when appropriate and tolerate them when necessary in a manner that fully permits reasoning to continue;
- d_3 make use of values beyond standard bivalence and standard trivalence (e.g., beyond e.g., Kleene’s, 1938 TRUE, FALSE, and UNKNOWN trio), specifically probabilities and strength factors (= cognitive likelihoods), (the latter case giving rise to multi-valued inductive logics);
- d_4 be argument-based, where the arguments have internal inference-to-inference structure, so that justification (and hence explanation) is available;
- d_5 have inference schemata (which sanction the inference-to-inference structure referred to in d_4), whether deductive or inductive, that are transparent, formal, and hence machine-checkable;
- d_6 be able to allow automated reasoning over the cognitive verbs/operators of knowledge, belief, desire, perception, intention, communication, etc., of the humans who are to be helped by this AI;
- d_7 be able to allow automated reasoning that can tackle Turing-unsolvable reasoning problems, e.g., queries about probability at and even above the *Entscheidungsproblem*. We do not here assume anything like hypercomputation. The requirement, here, is that formal science and engineering be harnessed to tackle *particular instances* of the Turing-uncomputable problem of algorithmically deciding provability.

We turn now to more detailed coverage of the technical background needed to understand our approach and its application in the promised three case studies.

6 Formal background of our brand of logicist AI

We first provide the reader with enough background to understand our approach and its application to the three case studies.

6.1 AI, logicist = logic-based AI, and artificial agents

AI has become a vast field as chronicled and explained in Bringsjord and Govindarajulu (2018). Accordingly, the pursuit of computing machines that qualify as intelligent and indeed even the meaning of “intelligent” itself in some contemporary debates are defined differently by different researchers and engineers, even

though all of them work under the umbrella of “AI.” Our approach is a logicist one, or—as it is sometimes said—a logic-based one. A full characterization of our approach to AI and robotics is of course beyond the reach of the present study, but we must give at least enough information to orient the reader and enable understanding of our three case studies, and we do so now. We turn first to the generic concept of an *artificial intelligent agent*, or—since, by context, it is clear that we must have intelligence, in some sense, front and center—simply *artificial agents*.

6.1.1 Artificial agents/AI, generically speaking

For present purposes, we rely upon how dominant textbooks, for example Russell and Norvig (2009, 2020); Luger (2008), characterize artificial agents. Their characterization is simply that such an agent computes a function from what is perceived (*percepts*) to behavior (*actions*). All such agents are assumed to operate this way in a certain *environment*, but for present purposes, we can leave explicit consideration of this aspect of the AI landscape to the side; doing so causes no loss of generality or applicability for the work we relate herein. However, what about the nature of the function from percepts to actions? As pointed out in the course of an attempt to show that the so-called Singularity²¹ is mathematically impossible (Bringsjord, 2012), the fact is that in the dominant AI textbooks, these functions are firmly assumed to be recursive. In the present study, we affirm this assumption, but the reader should keep in mind that despite this affirmation, our AI technology can still be based on automated reasoning that is routinely applied to problems that are Turing-uncomputable *in the general case*. This is directly expressed in desideratum d_7 in Des. After all, all automated reasoners that are specifically automated theorem provers for first-order logic confront the *Entscheidungsproblem*, first shown unsolvable by Church (Church’s Theorem). Our automated reasoners routinely attempt to discover arguments and proofs in order to settle queries at levels far above Church’s negative result.

6.1.2 The logicist approach to AI/robotics

We can now quickly state the heart of our logicist approach to AI and cognitive robotics as follows. The artificial agents we specify and implement compute their functions (from, again, percepts to actions) via automated reasoning over a given formula Φ in some formal language \mathcal{L} for some formal logic \mathcal{L} . This means that what these agents perceive must ultimately be transduced into content expressed in such formulae; and it means that an action, before translated into lower-level information that can trigger/control an effector, must also be expressed as a formula. The reader will see this in action below when we show our AI used in the trio of case studies. But how, specifically, are the functions computed in the case of such agents? The answer is straightforward: These functions are computed by automated reasoning. Of course, it has long been known that computation, while often understood in procedural

²¹ The point in future time at which, so the story goes, AIs reach human-level intelligence, and then immediately thereafter ascend to intellectual heights far, far above our own.

terms (e.g., in terms of Turing machines), is fully reducible to, and usable as, reasoning.²²

What about cognitive robotics, specifically? This is a key question because our Case Study 2 features our cognitive robot, PERI.2 (alert readers have noticed that we have already used the adjective “cognitive”). Alternatively, the introduction of cognitive elements to a formalism is said to make that formalism *behavioral* in nature; see Camerer, 2003.) We specifically pursue cognitive robotics as defined in the study by Levesque and Lakemeyer (2007),²³ with a slight formal tweak, and say simply that a cognitive robot is one whose macroscopic actions are a function of what the robot knows, believes, intends, and so on. As seen below, these verbs are at the heart of a *cognitive calculus*, the class of cognitively oriented logics we employ in general and in automated reasoning quite concretely. It will soon be observed that the robot PERI.2 is a cognitive robot, by the definitions just given and affirmed.

Our logicist-AI work is specifically enabled by *cognitive calculi*. Details regarding this class of logics and exactly how they are tailor-made for handling cognitive attitudes/verbs are provided in numerous publications in which such calculi are harnessed for various implementations (see Govindarajulu and Bringsjord, 2017a; Bringsjord et al., 2020b). Put with a brevity here that is sufficient, a cognitive calculus \mathcal{C} is a pair $\langle \mathcal{L}, \mathcal{I} \rangle$ where \mathcal{L} is a formal language (composed, in turn, minimally, of a formal grammar and an alphabet/symbol set), and \mathcal{I} is a collection of inference schemata (sometimes called a *proof theory* or *argument theory*) \mathcal{I} ; in this regard, our logicist-AI work is in the tradition of proof-theoretic semantics inaugurated by Prawitz (1972) and others (and for a modern treatment, see Francez, 2015; Bringsjord et al., 2022c).

Cognitive calculi have exclusively proof-theoretic and argument-theoretic semantics; no model theory is used, no possible worlds are used.²⁴ Within the present study, as explained below, dialects of the cognitive calculi *DCEC* (deductive) and *IDCEC* (inductive) will be utilized, and this is what makes success in our case studies in Section 7 possible.

We said that *IDCEC* is an inductive cognitive calculus. The great pioneer of modern inductive logic in any form was Rudolph Carnap. Carnap would say, and the logicians and mathematicians today who continue his particular approach in the form of what is known as *pure inductive logic* (PIL) (Paris and Vencovská, 2015) would still say, that “There is a blue cube on the table” and “There is an orange sphere on the table” should each be assigned

a probability value (a real number between 0 and 1, inclusive), and this content, combined with additional probabilized propositions, can be used in a process that dictates what should be rationally believed. Unfortunately, Carnap and his followers pay precious little attention to the “coin of the realm” in human reasoning and decision-making: arguments and argumentation. This tradition (which began long before Carnap and includes e.g., Keynes and Bayes) also runs afoul of the brute fact that humans very rarely use probabilities and the probability calculus. In our approach, to computational inductive logic for AI, inference schemata that, when instantiated in sequence, lead to arguments and proofs, are front and center. This can be observed clearly in the specifications of both of the cognitive calculi used in the present study, which we now provide (next section). Later, in the three forthcoming case studies, it is the automated discovery of arguments and proofs based on linked inferences as instantiations of these schemata that is key.

6.2 Cognitive calculi, in more detail

Cognitive calculi, as we have said, are members of an infinite family of highly expressive logics that, for instance, include unrestricted third-order logic, meta-logical quantification, and predication (it can be expressed not only that a property has a property but that a formulae has a property), and all this extensional machinery is intertwined with intensional operators for belief, knowledge, intention, communication, action, and the traditional alethic modalities as well. To the best of our knowledge, cognitive calculi are the most expressive logics that have been implemented and used with corresponding automated reasoners. For more on cognitive calculi, see Arkoudas and Bringsjord (2009a); Govindarajulu and Bringsjord (2017a); Govindarajulu et al. (2019); Bringsjord et al. (2020b). For the shortest account of cognitive calculi, and implementation of reasoning over declarative content therein, in which it is made clear that such calculi are exclusively proof- and argument-theoretic, see Bringsjord and Govindarajulu (2020). For an explanation of how natural-language understanding works in connection with cognitive calculi, see Bringsjord et al. (2022c). There are many more resources available, as cognitive calculi are well established at this point, but for present purposes, it suffices to economically provide the specifications of the two cognitive calculi used for modeling and simulation in the present study, and these specifications follow now.

6.2.1 Specifications of cognitive calculi *DCEC* and *IDCEC*

Below is the signature of the standard dialect of *DCEC*. The signature contains the sorts, function signatures, and grammar of this cognitive calculus, presented in a manner that is standard and self-explanatory for the most part. As obvious, lower-case Greek letters are formulae, bolded majuscule Roman letters are intensional/modal operators (**K** for *knows*, **B** for *believes*, **I** for *intends*, etc.).

²² This is what allows proofs of the Halting Problem for Turing machines to be relied upon to prove the undecidability of the *Entscheidungsproblem*; see Boolos et al. (2003).

²³ As pointed out in that study, as far as most relevant thinkers know, it was actually Ray Reiter (the same thinker who introduced default logic, briefly mentioned above) who coined and first defined the phrase “cognitive robotics.”

²⁴ Bringsjord’s rejection of possible-worlds semantics can be traced to his proof rather long ago that such structures can be shown to be mathematically impossible; see Bringsjord (1985).

DCEC Signature

$S ::= \text{Agent} \mid \text{ActionType} \mid \text{Action} \sqsubseteq \text{Event} \mid \text{Moment} \mid \text{Fluent}$

$f ::= \begin{cases} \text{action} : \text{Agent} \times \text{ActionType} \rightarrow \text{Action} \\ \text{initially} : \text{Fluent} \rightarrow \text{Formula} \\ \text{holds} : \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \text{happens} : \text{Event} \times \text{Moment} \rightarrow \text{Formula} \\ \text{clipped} : \text{Moment} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \text{initiates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \text{terminates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \text{prior} : \text{Moment} \times \text{Moment} \rightarrow \text{Formula} \end{cases}$

$t ::= x : S \mid c : S \mid f(t_1, \dots, t_n)$

$\phi ::= \begin{cases} q : \text{Formula} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \forall x : \phi(x) \mid \exists x : \phi(x) \\ \mathbf{P}(a, t, \phi) \mid \mathbf{K}(a, t, \phi) \mid \mathbf{S}(a, b, t, \phi) \mid \mathbf{S}(a, t, \phi) \\ \mathbf{C}(t, \phi) \mid \mathbf{B}(a, t, \phi) \mid \mathbf{D}(a, t, \phi) \mid \mathbf{I}(a, t, \phi) \\ \mathbf{O}(a, t, \phi, \neg)\text{happens}(\text{action}(a^*, \alpha), t') \\ \text{Perceives, Knows, Says, Common-knowledge} \\ \text{Believes, Desires, Intends, Ought-to} \end{cases}$

Next is the standard set of inference schemata for *DCEC*. They say that when what is above the vertical line is instantiated, that which is below can be inferred (in accordance with that instantiation); this top-bottom notation is common in descriptions of so-called *natural deduction*. The approach to logicist AI-based on cognitive calculi is not restricted in any way to “off the shelf” logics but are instead created and specified for given purposes and applications in AI. However, all cognitive calculi include standard extensional logics (one or more of $\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$, and standard natural-inference schemata for these extensional logics).

DCEC Inference Schemata

$\frac{\mathbf{K}(a, t_1, \Gamma), \Gamma \vdash \phi, t_1 \leq t_2}{\mathbf{K}(a, t_2, \phi)} [I_K] \quad \frac{\mathbf{B}(a, t_1, \Gamma), \Gamma \vdash \phi, t_1 \leq t_2}{\mathbf{B}(a, t_2, \phi)} [I_B]$

$\frac{}{\mathbf{C}(t, \mathbf{P}(a, t, \phi) \rightarrow \mathbf{K}(a, t, \phi))} [I_1] \quad \frac{}{\mathbf{C}(t, \mathbf{K}(a, t, \phi) \rightarrow \mathbf{B}(a, t, \phi))} [I_2]$

$\frac{\mathbf{C}(t, \phi), t \leq t_1, \dots, t \leq t_n}{\mathbf{K}(a_1, t_1, \dots, \mathbf{K}(a_n, t_n, \phi) \dots)} [I_3] \quad \frac{\mathbf{K}(a, t, \phi)}{\phi} [I_4]$

$\frac{t_1 \leq t_2 \leq t_3}{\mathbf{C}(t, \mathbf{K}(a, t_1, \phi_1) \rightarrow \phi_2) \rightarrow \mathbf{K}(a, t_2, \phi_1) \rightarrow \mathbf{K}(a, t_3, \phi_2)} [I_5]$

$\frac{t_1 \leq t_2 \leq t_3}{\mathbf{C}(t, \mathbf{B}(a, t_1, \phi_1) \rightarrow \phi_2) \rightarrow \mathbf{B}(a, t_2, \phi_1) \rightarrow \mathbf{B}(a, t_3, \phi_2)} [I_6]$

$\frac{t_1 \leq t_2 \leq t_3}{\mathbf{C}(t, \mathbf{C}(t_1, \phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{C}(t_2, \phi_1) \rightarrow \mathbf{C}(t_3, \phi_2)} [I_7]$

$\frac{}{\mathbf{C}(t, \forall x. \phi \rightarrow \phi[x \mapsto t])} [I_8] \quad \frac{}{\mathbf{C}(t, \phi_1 \leftrightarrow \phi_2 \rightarrow \neg\phi_2 \rightarrow \neg\phi_1)} [I_9]$

$\frac{}{\mathbf{C}(t, [\phi_1 \wedge \dots \wedge \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \dots \rightarrow \phi_n \rightarrow \phi])} [I_{10}]$

$\frac{\mathbf{B}(a, t, \phi) \quad \mathbf{B}(a, t, \phi \rightarrow \psi)}{\mathbf{B}(a, t, \psi)} [I_{11a}] \quad \frac{\mathbf{B}(a, t, \phi) \quad \mathbf{B}(a, t, \psi)}{\mathbf{B}(a, t, \phi \wedge \psi)} [I_{11b}]$

$\frac{\mathbf{S}(s, h, t, \phi)}{\mathbf{B}(h, t, \mathbf{B}(s, t, \phi))} [I_{12}] \quad \frac{\mathbf{I}(a, t, \text{happens}(\text{action}(a^*, \alpha), t'))}{\mathbf{P}(a, t, \text{happens}(\text{action}(a^*, \alpha), t'))} [I_{13}]$

$\frac{\mathbf{B}(a, t, \phi) \quad \mathbf{B}(a, t, \mathbf{O}(a, t, \phi, \chi))}{\mathbf{K}(a, t, \mathbf{I}(a, t, \chi))} [I_{14}]$

The following two framed boxes specify the additional signature and inference schemata for *IDCEC*, respectively. That is, they build on top of those given for *DCEC* immediately above. These specifications enable reasoning about uncertain belief. In the first of three case studies discussed next, we will describe the uncertainty system which enables the ascription of *likelihood* values to beliefs present in these schemata. Herein, we only

provide a subset of the inference schemata of *IDCEC*; a full exposition of *IDCEC* and its inference schemata are the focus of a doctoral dissertation (Giancola, 2023). For an early inductive cognitive calculus with cognitive likelihood, see Govindarajulu and Bringsjord (2017b).

Additional Signature for IDCEC

$S ::= \text{Number} \mid \text{List}$

$f ::= \begin{cases} \text{min} : \text{List}[\text{Number}] \rightarrow \text{Number} \\ \text{max} : \text{List}[\text{Number}] \rightarrow \text{Number} \end{cases}$

$\phi ::= \begin{cases} \mathbf{B}^\sigma(a, t, \phi) \end{cases}$

where $\sigma \in [-5, -4, \dots, 4, 5]$

Additional Inference Schemata for IDCEC

$\frac{\mathbf{S}(s, a, t_1, \phi), t_1 < t_2}{\mathbf{B}^1(a, t_2, \phi)} [I_1^c] \quad \frac{\mathbf{P}(a, t, \phi)}{\mathbf{B}^4(a, t, \phi)} [I_4^c]$

$\frac{\mathbf{B}^\sigma(a, t_1, \phi), \Gamma \not\vdash \neg\mathbf{B}^\sigma(a, t_2, \phi), t_1 < t_2}{\mathbf{B}^\sigma(a, t_2, \phi)} [I_{PROP}^\ell]$

$\frac{\mathbf{B}^{\sigma_1}(a, t, \phi_1), \dots, \mathbf{B}^{\sigma_m}(a, t, \phi_m), \{\phi_1, \dots, \phi_m\} \vdash \phi, \{\phi_1, \dots, \phi_m\} \not\vdash \perp}{\mathbf{B}^{\min(\sigma_1, \dots, \sigma_m)}(a, t, \phi)} [I_{WLP}^\ell]$

where $\sigma_i \in [0, 1, \dots, 4, 5]$

6.2.2 Regarding metatheoretical properties of our cognitive calculi and associated automated reasoners

As the chief purpose of the study we report herein is to advance logicist AI, both formally and computationally, rather than to advance computational formal logic in and of itself, it would be inappropriate to spend appreciable time and space explaining, let alone proving, the metatheoretical properties—soundness, completeness, un/decidability, complexity measures, etc.—of the family of cognitive calculi and the members thereof used herein (*DCEC* & *IDCEC*) and of our automated reasoners. However, we do now provide some brief metatheoretical information that readers well versed in formal logic will likely find helpful.

To begin, recall that desideratum d_7 , if satisfied, ensures that the fundamental question as to whether some formula ϕ can be inferred (*via* some collection of inference schemata) from some set Φ of formulae is for us usually²⁵ Turing-undecidable. We have already mentioned Church’s Theorem in this regard, which of course applied to theoremhood in first-order logic = \mathcal{L}_1 . However, as a matter of fact, \mathcal{L}_1 is *semi-decidable*: if, in fact, there exists a

²⁵ There will be the off case of a query, e.g., as to whether a low-expressivity ϕ is inferable from a low-expressivity Φ , for instance when all formulae selected for automated processing are mere propositional-calculus formulae, or—more realistically—when all formulae fall into a decidable fragment of \mathcal{L}_1 , e.g., fluted logic. However, the standard cases for use of cognitive calculi, which are multi-modal quantified logics, will include high-expressivity formulae.

proof in the first-order case that supports an affirmative answer to the question, that proof can be algorithmically found. However, in the case of our paradigm, there are many general inference questions posable by and to our artificial agents using as a basis a cognitive calculus (whether deductive or inductive) that are fully undecidable. This can be immediately observed from the well-known theorem that \mathcal{L}_2 is not even semi-decidable.²⁶ However, our study, as it is based on cognitive calculi, places crucial reliance upon human-level cognitive verbs, where these verbs are logicized by relevant modal operators; for example: **P** for *perceives*, **B** for *believes* (which, in our approach, can have a positive likelihood parameter attached), **K** for *knows* (which also can have a positive likelihood parameter attached), and so on. This means that things are only that much harder computationally, and in fact, since both the Arithmetic and Analytic Hierarchies are purely extensional (the former based on \mathcal{L}_1 and the latter based on \mathcal{L}_2), and hence devoid of modal operators, things are only even harder, given our willingness to consider formulae and queries arising from an unflinching look at the human case. This is simply the nature of the beast—that beast being the undeniable expressivity of human-level cognition and specifically of human-level argumentation. After all, there can be no denying that humans create and assess arguments that, when logicized, require remarkably high levels of expressivity; this holds for even everyday activity, not just for *recherché* academic problems. For an everyday example, let us consider an argument, to be found and verified by our AI technology, for the proposition (\ddagger) that the dog Rover is scary, based chiefly on these two premises:

- (P1) As trainer David knows, there are some properties that are downright scary and that some dogs have; and if they have any of these properties, the dog in question is itself scary.
- (P2) David also knows that one of these scary properties is having prominent and pronounced musculature, and another is having long and large incisors.

Now further suppose that (P3) David perceives a particular dog, Rover, who as it happens has thick, pronounced incisors and prominent pronounced musculature. Our automated reasoner, ShadowProver, working with the formal representation of $\{P1, P2, P3\}$ in the cognitive calculus \mathcal{DCEC}^3 ,²⁷ is able to find an argument, and verify it, for (\ddagger)—despite the formal fact that, in the general case, the question as to whether a proposition follows from modalized third-order formulae is a Turing-undecidable question.²⁸

Some readers, even cognoscenti, may then ask: But if the queries your artificial agents much seek to handle are this difficult, how does the engineering of your automated-reasoning systems work? This question alone, if answered fully, would require its own monograph. However, the answer is actually quite simple,

fundamentally, The short version of the answer is that our engineering (a) reflects the famous conception, originated by AI pioneer Herbert Simon, of “satisficing” (Simon, 1956); and (b) this engineering makes use of a most valuable but low-technology subsystem: a stopwatch, in the form of timeouts on duration of CPU processing. In other words, we engineer for success on particular cases within the general space of Turing-uncomputable problems, and if processing takes too long and no answer has been returned, we curtail processing by fiat, in accordance with a pre-set length of time allowed for CPU activity. In the case of our three case studies featured herein, temporal thresholds were not reached, in fact were not even approached.²⁹

What about other metatheoretical properties in the realm of formal logic? What about complexity, soundness, completeness, for example? Complexity is irrelevant, because almost all of the problems that our human- and argumentation-centric artificial agents seek to solve are not even in the Polynomial Hierarchy (since they are above Σ_1 in the Arithmetic Hierarchy). Soundness and completeness, given that our approach is purely proof-theoretic, is beyond scope; readers for a start are directed to Govindarajulu et al. (2019). As can be readily understood given the foregoing, while there is a lot of truly impressive work in AI and intelligent systems that makes use of computational logic, much of it is nonetheless radically different in formal orientation than ours. An example is the use of logic programming. For a specific example, as Brewka et al. (2011) show, *answer set programming* (ASP) is quite powerful and promising—but its nature is applauded and affirmed because “ASP . . . aim[s] to maintain a balance between expressivity, ease of use, and computational effectiveness” (Brewka et al., 2011, p. 92–93). The balance, here, can indeed be very powerful, but as should be abundantly clear, our approach and the concrete case studies within it reported herein, we do not desire this balance.³⁰

One final word, aimed especially at those who subscribe, as the first author long did but no longer does, to the general expressivity-vs.-tractability tradeoff for formal (extensional) logics that has become part of the fixed furniture of logicist AI. This tradeoff, entrenched since at least the publication of the important (Levesque and Brachman, 1985), is far from being both clear and ironclad in the case of our brand of AI engineering. The logico-mathematical reason stems directly from Gödel’s Speedup Theorem (GST) (Buss, 1994, 1995), which, in word, says that the move from first- to-second-order logic enables a non-recursive gain in efficiency, measured by length of proof (and likewise for jumping from second- to-third-order, and so on for each jump).³¹ In engineering terms, while of course we have no recourse to algorithms for answering queries fully in the general case, we also know that engineering techniques just might find staggering gains in efficiency for cases at hand. Readers interested in learning

26 In general, once one moves beyond first-order logic, a dramatic loss of metatheoretical properties desirable to many (not us) occurs, as revealed in Lindström’s Theorems, elegantly covered in Ebbinghaus et al. (1994).

27 When the extensional core is \mathcal{L}_3 .

28 We do not spend the space to recount why.

29 In fact, every run arising from every query that triggers automated reasoning/planning in our three case studies is clocked in milliseconds: no run exceeded 3 s on an Apple laptop.

30 We do not by the way mean to imply that no one within the ASP rubric has tackled human-level cognition. Ganascia (2007), e.g., has modeled and simulated aspects of lying constrained by this rubric.

31 Gödel’s results pertain directly only to elementary number theory, but they carry over their application to other domains.

more about this phenomenon are advised to start with the striking example of Boolos (1987) and move from there to study GST itself via the references we provided.

7 Three Case Studies

We turn now to our three case studies. In the third and final study, reasoning is explained in somewhat higher-level terms than in the case of the first and second; more specifically, the arguments in Case Study 3 are for space-saving and expository purposes expressed rather informally. Our first study takes us back to Monty Hall, and we proceed to it now.

7.1 Case Study 1: MHP₃ redux

We have every confidence the reader will remember MHP₃, which we suppose that some artificial agents have perceived in full, automatically discovered the correct answer for, and now share that answer with a typical human who fails to grasp the problem, and thought the correct answer was STAY. How helpful is this artificial agent going to be to this human? Not very. After all, our human does not know *why* the correct answer is SWITCH. The obvious solution, given the need for genuinely helpful human-centric AI, is a class of artificial agents that can not only find solutions but also provide cogent, compelling, verified arguments, certifying those solutions. If such a thing is provided in the present case, the human will be enlightened. This is what our artificial agents can do.

Given the complexity of MHP₃, we cannot, herein, canvass the full terrain of this problem, its logicization into our inductive logic *IDCEC*, and solutions automatically found, but let us consider two prominent arguments regarding MHP₃, the first sound (and hence both veracious and valid³²) and the second not. The sound argument goes as follows:

1. Without loss of generality, assume that you select Door 1.³³
2. There are three potential cases, in which the prize is behind Door 1, Door 2, or Door 3, respectively.
3. Let's first consider the outcome of the three cases under the STAY protocol.
 - (a) If the prize is behind Door 1, you win. If it is behind Door 2 or 3, you lose.
 - (b) Hence there is a $\frac{1}{3}$ chance of winning if you follow STAY.
4. The cases are a bit more complex if you follow SWITCH, because, crucially, Monty *knows* where the prize is, and, having *perceived* your initial choice, will *always* reveal a door without the prize behind it.

³² Following long-established and customary terminology, a *sound* argument is both *veracious* and *valid*; a *veracious* argument has true (or at least plausible to some level in some stratification of plausibility) premises; and a *valid* argument has inferences that abide by the collection of inference schemata taken to be operative in the case at hand.

³³ Since it is equally likely that the prize is behind any of the three doors, the same argument can be generated regardless of which door is initially selected.

- (a) If the prize is behind Door 1, you will lose. Monty can open either of Door 2 or Door 3 (and should be assumed to randomly choose which one), and regardless of which door you switch to, you will lose.
- (b) If the prize is behind Door 2, Monty *must* open Door 3. Therefore if you SWITCH to Door 2, you will win.
- (c) If the prize is behind Door 3, Monty *must* open Door 2. Therefore, if you follow SWITCH and move to Door 2, you will win.
- (d) Hence, by simply counting, we deduce that there is a $\frac{2}{3}$ chance of winning if you follow SWITCH.

While many arguments have been made for STAY,³⁴ they mostly follow the same general pattern. That pattern is as follows:

1. Without loss of generality, assume that you select Door 1, and that Monty then opens Door 3.
2. When Monty opens Door 3 that door of course has dropped out of consideration, and we are down to two doors, so the probability that the prize is behind Door 1 becomes $\frac{1}{2}$; same as the probability that the prize is behind Door 2.
3. Hence there is no reason to switch doors (and since—as the economists who study rationality say—time is money, switching is irrational).

Pinpointing where this invalid argument goes awry is enabled by our concept of *likelihood*, specifically what we term *cognitive likelihood* (Giancola, 2023). The invention of this concept and its use in our intelligent, defeasible argumentation systems satisfies desideratum *d*₃. This concept enables the ranking of the strength of beliefs (and other cognitive attitudes), in accordance with their likelihood values. The spectrum of the 11 possible values are presented in Table 1 (the caption for which offers some contextualization of these values in contrast with probabilities). The use of these strength-factor/cognitive likelihood values makes *IDCEC* a multi-valued (or many-valued) logic; an efficient, broad overview of such logics is provided in the study by Gottwald (2015).³⁵

By enabling beliefs to take on these uncertainty levels, cognitive likelihood allows agents to reason with uncertain beliefs generated by and reasoned over in integration with other modalities, for example, with perception, communication, and intention. This is formalized in the inference schemata of *IDCEC*. For example,

³⁴ See Pinker (2021) for an argument (authored and advocated by a mathematician) made by a analogy to horse race, and for more extensive coverage of such arguments, see Granberg (2014). We explain in the study mentioned in Section 8 that two-horse arguments are ideal specimens of sophisticated argumentation.

³⁵ To the best of our knowledge, while the first multi-valued modal logic (a three-valued one) appeared in 1967 due to the study by Segerberg (1967), *IDCEC* is the first multi-valued *multi-modal-operator* logic, and with little question (for better or worse), the first such logic computationally implemented. Multi-valued logics in the non-modal/extensional case (three values) originated with the study by Łukasiewicz (1920), and our basis on the extensional side (e.g., standard \mathcal{L}_1) for our cognitive calculi is an extension and refinement of Łukasiewicz's study by Kleene's (1938); see again desideratum *d*₃ in Des.

TABLE 1 The 11 cognitive likelihood values.

Numerical	Linguistic
5	CERTAIN
4	EVIDENT
3	OVERWHELMINGLY LIKELY
	= BEYOND REASONABLE DOUBT
2	LIKELY
1	MORE LIKELY THAN NOT
0	COUNTERBALANCED
-1	MORE UNLIKELY THAN NOT
-2	UNLIKELY
-3	OVERWHELMINGLY UNLIKELY
	= BEYOND REASONABLE BELIEF
-4	EVIDENTLY NOT
-5	CERTAINLY NOT

These values, notably, are not in any way real numbers in an interval, as are probabilities in Kolmogorov's (1933) probability calculus (the interval of course being $[0, 1]$), much used in modern AI, e.g., in Bayesian approaches. Rather, these are fixed values in the traditional sense of 'value' in multi-valued (or many-valued) logics, where each value has an independent justification as a determinate value in rational human cognition. For example, when strength/value is 3 for a belief, this corresponds to what humans in general refer to as something that ought (epistemically, not morally, speaking) to be believed because the proposition is "beyond reasonable doubt," a concept central to occidental jurisprudence. For the present study, it is beyond scope to present our full axiomatic theory \mathcal{L} of cognitive likelihood that is subsumed by \mathcal{IDCEC} , in which Kolmogorov's axioms do not hold. E.g., where p yields the probability of an event/proposition ϕ , Kolmogorov's second axiom says that if ϕ is a theorem in a standard, elementary extensional logic (such as the propositional calculus), $p(\phi)=1$. However, theorems in such a logic are not at all guaranteed to have a likelihood value of 5, since an infinite number of such theorems are not familiar to human beings and hence cannot be believed. In addition, theorems of \mathcal{L} are often completely without corresponding analogs in the probability calculus. E.g., "if $\ell(\phi) = 5$, $\ell(\neg\phi) = 0$ " is a theorem in \mathcal{L} that has no analog in the probability calculus.

perception of ϕ sanctions, by inference schema I_4^ℓ (see the specification of inference schemata in the specifications shown in Section 6.2.1), a belief that ϕ —but only at the cognitive-likelihood value $\sigma := 4$. (that which we perceive, at least when we are talking about perception of things in the external world, might be illusory). Certainty, when $\sigma := 5$, is reserved in our framework for belief regarding mathematical propositions. In general, this ability to reason with cognitive-likelihood values enables the kind of nuanced argumentation we seek, as it provides a formalism in which individual statements and arguments as a whole can be assigned relative strengths (= cognitive likelihoods), which, in turn, allows certain statements and arguments carrying higher strength to "defeat" others non-monotonically as time flows; this occurs in our case studies.

Now, back to MHP₃. The first argument is fully supported by the basic tenets of probability theory viewed through the lens of odds (i.e., the probability of an event is the ratio of the number of possible outcomes in which it occurs, over the number of total possible outcomes).³⁶ Therefore, a belief in

the conclusion of Argument 1—namely, that one should follow SWITCH—can be held at the level of EVIDENT. It is EVIDENT, not CERTAIN, because the argument fundamentally relies on the agent's perception of various elements of the game, which could be compromised without violation of any mathematically necessary axioms or theorems. Such beliefs are inferred using schema I_4^ℓ as follows:

$$\frac{P(a, t, \phi)}{B^4(a, t, \phi)} [I_4^\ell] \quad (1)$$

On the other hand, Step 2. of Argument 2 is generally asserted with no justification. One could argue that it is justified by the large group of people who state it. Given the inference schema $[I_2^\ell]$, such a justification can warrant a belief at the level of MORE LIKELY THAN NOT but not higher. Therefore, we have formally observed that the first argument is stronger than the other and hence should be accepted.

As mentioned above, while a full formal and computational account of the overarching argument and its sub-proofs are out of scope in the present study, we give the automated proofs found by ShadowAdjudicator in Figure 1 and point the interested reader to Giancola (2023) for a full exposition of the relevant inference schemata, all the arguments and proofs, and full analysis. We mention as well that there are now numerous variants of MHP₃ that are a good deal trickier than the original; these are comprehensively treated in the study by Bringsjord et al. (2022b), which takes account, for instance, of the variants discussed in the study by Rosenthal (2008).

7.2 Case Study 2: the robot PERI.2 meets "Clouded" Meta-Forms

Our second case study revolves around a very interesting and challenging reasoning game that we are using in a sustained attempt to quite literally have the cognitive robot PERI.2³⁷ attend school and progress grade-by-grade through at least high school, on the road thereby to artificial general intelligence (AGI); this project was announced in Bringsjord et al. (2022a). The game is called "Meta-Forms" (see Figure 2 for a rapid orientation to the game).

For our second case study, PERI.2 is issued the challenge of solving a Meta-Forms problem; not one of the very hardest of such problems, but certainly a non-trivial one, even for adult humans; the problem is shown in Figure 3.

PERI.2 does meet with success, in what as far as we know is one of the most robust uses of argumentation-based AI in cognitive robotics. This success is shown in Figure 4, and the automatically found reasoning that leads to PERI.2's knowledge³⁸ (which, in turn, leads to the intention to act accordingly, and then the performance of

cognitive calculi that subsume the two— $DCEC$ and $IDCEC$ —we employ herein, but this is out of scope.

³⁷ The precursor robot, PERI, anchored the introduction, to the field of AI, what is called *psychometric AI*; see Bringsjord and Schimanski (2003).

³⁶ This approach to probability can be formalized in what is known as *probability logic* (Adams, 1998), and probability logic can be subsumed in

```
(base) root@97e884a1add6:/base# python diss_examples/new_monty_hall_problem.py
Modeling Valid Reasoning in MHP...
Reasoning About Outcomes if the Prize is Behind Door 2...
set to 2
(If Intro
  (FOLFromSnark
    Givens:
      (Implies (and (Selects c1 t3 d1) (CarBehind d1)) win)
      (Knows!_monty_t1_CarBehind_d2_)
      (Implies (and (Opens monty t2 d2) switch) (Selects c1 t3 d3))
      (Knows!_c1_t2_implies_and_Opens_monty_t2_d3_stay__Selects_c1_t3_d1_)
      (Implies (and (Opens monty t2 d3) switch) (Selects c1 t3 d2))
      (Believes!_c1_t2_implies_and_Opens_monty_t2_d3_stay__Selects_c1_t3_d1_)
    (switch
      (GIVEN{}))
    ((Selects c1 t1 d1)
      (GIVEN{}))
      (Believes!_c1_t2_implies_and_Opens_monty_t2_d2_switch__Selects_c1_t3_d3_)
      (Implies (and (Selects c1 t1 d1) (CarBehind d3)) (Opens monty t2 d2))
      (Believes!_c1_t2_implies_and_Opens_monty_t2_d3_switch__Selects_c1_t3_d2_)
      (Implies (and (Selects c1 t1 d1) (CarBehind d1)) (Opens monty t2 d2))
      (Knows!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d2__Opens_monty_t2_d3_)
      (Knows!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d1__Opens_monty_t2_d2_)
      (Implies (and (Opens monty t2 d3) stay) (Selects c1 t3 d1))
      (Knows!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d3__Opens_monty_t2_d2_)
      (Implies (and (Selects c1 t3 d2) (CarBehind d2)) win)
      (Believes!_c1_t2_implies_and_Opens_monty_t2_d2_switch__Selects_c1_t3_d3_)
      (Knows!_monty_t1_CarBehind_d2_)
      (Believes!_monty_t1_CarBehind_d2_)
      (Knows!_c1_t2_implies_and_Opens_monty_t2_d3_switch__Selects_c1_t3_d2_)
      (forall (?d) (Implies (and (Selects c1 t3 ?d) (CarBehind ?d)) win))
      (Believes!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d2__Opens_monty_t2_d3_)
    )
    ((CarBehind d2)
      (GIVEN{}))
      (Implies (and (Selects c1 t1 d1) (CarBehind d2)) (Opens monty t2 d3))
      (Knows!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d1__Opens_monty_t2_d2_)
      (Believes!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d3__Opens_monty_t2_d2_)
      (Implies (and (Selects c1 t3 d3) (CarBehind d3)) win)
    )
    Goals:
      win)))
PROOF OF: (Believes!_l=2,p=2/3) c1 t1 (implies switch win))
Applied 'Probabilistic Belief Intro' to: (Believes!4 c1 t1 (Odds! (implies switch win) (POS case2 case3) (NEG case1)))
PROOF OF: (Believes!4 c1 t1 (Odds! (implies switch win) (POS case2 case3) (NEG case1)))
GIVEN
```

```
(base) root@97e884a1add6:/base# python diss_examples/new_monty_hall_problem.py
Reasoning About Outcomes if the Prize is Behind Door 1...
(If Intro
  (FOLFromSnark
    Givens:
      (Believes!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d1__Opens_monty_t2_d3_)
      (Implies (and (Selects c1 t3 d1) (CarBehind d1)) win)
      (Implies (and (Opens monty t2 d2) switch) (Selects c1 t3 d3))
      (Believes!_monty_t1_CarBehind_d1_)
      (Knows!_c1_t2_implies_and_Opens_monty_t2_d3_stay__Selects_c1_t3_d1_)
      (Implies (and (Opens monty t2 d3) switch) (Selects c1 t3 d2))
      (Believes!_c1_t2_implies_and_Opens_monty_t2_d3_stay__Selects_c1_t3_d1_)
    ((CarBehind d1)
      (GIVEN{}))
      ((Selects c1 t1 d1)
        (GIVEN{}))
        (Believes!_c1_t2_implies_and_Opens_monty_t2_d2_switch__Selects_c1_t3_d3_)
        (Implies (and (Selects c1 t1 d1) (CarBehind d3)) (Opens monty t2 d2))
        (Believes!_c1_t2_implies_and_Opens_monty_t2_d3_switch__Selects_c1_t3_d2_)
        (Knows!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d2__Opens_monty_t2_d3_)
        (Implies (and (Opens monty t2 d3) stay) (Selects c1 t3 d1))
        (Knows!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d3__Opens_monty_t2_d2_)
        (Implies (and (Selects c1 t3 d2) (CarBehind d2)) win)
        (Knows!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d1__Opens_monty_t2_d3_)
        (Knows!_c1_t2_implies_and_Opens_monty_t2_d2_switch__Selects_c1_t3_d3_)
        (Knows!_c1_t2_implies_and_Opens_monty_t2_d3_switch__Selects_c1_t3_d2_)
        (forall (?d) (Implies (and (Selects c1 t3 ?d) (CarBehind ?d)) win))
      )
      (stay
        (GIVEN{}))
        (Believes!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d2__Opens_monty_t2_d3_)
        (Knows!_monty_t1_CarBehind_d1_)
        (Implies (and (Selects c1 t1 d1) (CarBehind d2)) (Opens monty t2 d3))
        (Implies (and (Selects c1 t1 d1) (CarBehind d1)) (Opens monty t2 d3))
        (Believes!_monty_t1_implies_and_Selects_c1_t1_d1_CarBehind_d3__Opens_monty_t2_d2_)
        (Implies (and (Selects c1 t3 d3) (CarBehind d3)) win)
      )
      Goals:
        win)))
PROOF OF: (Believes!_l=2,p=1/3) c1 t1 (implies stay win))
Applied 'Probabilistic Belief Intro' to: (Believes!4 c1 t1 (Odds! (implies stay win) (POS case1) (NEG case2 case3)))
PROOF OF: (Believes!4 c1 t1 (Odds! (implies stay win) (POS case1) (NEG case2 case3)))
GIVEN
```

Modeling Invalid Reasoning in MHP...

```
PROOF OF: (and (Believes!_l=2,p=1/2) c2 t2 (CarBehind d1)) (Believes!_l=2,p=1/2) c2 t2 (CarBehind d2)))
Applied 'Modus Ponens' to: (implies (Believes!4 c2 t2 (not (CarBehind d3))) (and (Believes!_l=2,p=1/2) c2 t2 (CarBehind d1)) (Believes!_l=2,p=1/2) c2 t2 (CarBehind d2)))) (Believes!4 c2 t2 (not (CarBehind d3)))
PROOF OF: (implies (Believes!4 c2 t2 (not (CarBehind d3))) (and (Believes!_l=2,p=1/2) c2 t2 (CarBehind d1)) (Believes!_l=2,p=1/2) c2 t2 (CarBehind d2))))
GIVEN
PROOF OF: (Believes!4 c2 t2 (not (CarBehind d3)))
Applied 'I'vell_4' to: (Perceives! c2 t2 (not (CarBehind d3)))
PROOF OF: (Perceives! c2 t2 (not (CarBehind d3)))
GIVEN
```

FIGURE 1

Two arguments for supposedly solving MHP₃, automatically found by ShadowAdjudicator/ShadowProver. The complete valid argument includes six sub-proofs, the result of considering whether switching or staying will result in a win depending on the three possible locations of the prize (and assuming, without loss of generality, that the contestant initially selected Door 1). In the graphic here, we show two of the six: switching when the prize is behind Door 2, and staying when the prize is behind Door 1. One of the others is the same as one shown: the contestant wins if they switch when the prize is behind Doors 2 or 3. The other 3 proofs result in failure; e.g., one cannot prove that staying will result in a win if the prize is behind Doors 2 or 3.

the action) is shown in Figure 5. It is important to realize that because of the nature of Meta-Forms problems, dynamic argumentation through time is part and parcel of how PERI.2 operates.

However, now what happens if PERI.2's environment is uncooperative? Specifically, what happens when this cognitive robot is faced with fog (or smoke, etc.), to the point where some possibly crucial information cannot be perceived, then believed, and then reasoned about? Such a situation is shown in Figure 6. In this situation, PERI.2 is unable to arrive at knowledge in support of action that can be taken in order to physically solve the problem (see Figure 7).

38 In the case of the step presented in Figure 4, PERI.2 is able to utilize disjunctive syllogism to satisfy the probability query in schema $[I_k]$. Essentially, because PERI.2 knows that there are already puzzle pieces in three of the four possible places it can put the blue piece, the piece must go in the only remaining place.

7.3 Case Study 3: a life-and-death multi-agent decision

The ARCADIA human-level cognitive architecture (Bridewell and Bello, 2015) provides means by which we are able to integrate our cognitive calculi and associated automated reasoners with a perceptual system that takes into account not only the general cognitive science of perception but also specifically a given agent's dynamically shifting attention. Computational cognitive science has disclosed that attention and perception go hand in symbiotic hand, and when an agent is designed and implemented as an ARCADIA model, this symbiosis is made computationally real.

In the present section, we give a case study of a robust multi-agent system perceiving and reasoning, and in which our automated-reasoning technology helps assess threat levels in a delicate scenario that is too depressingly real in the world today. The simulation is in real time, as perceptual information is communicated to and from multiple agents. However, before

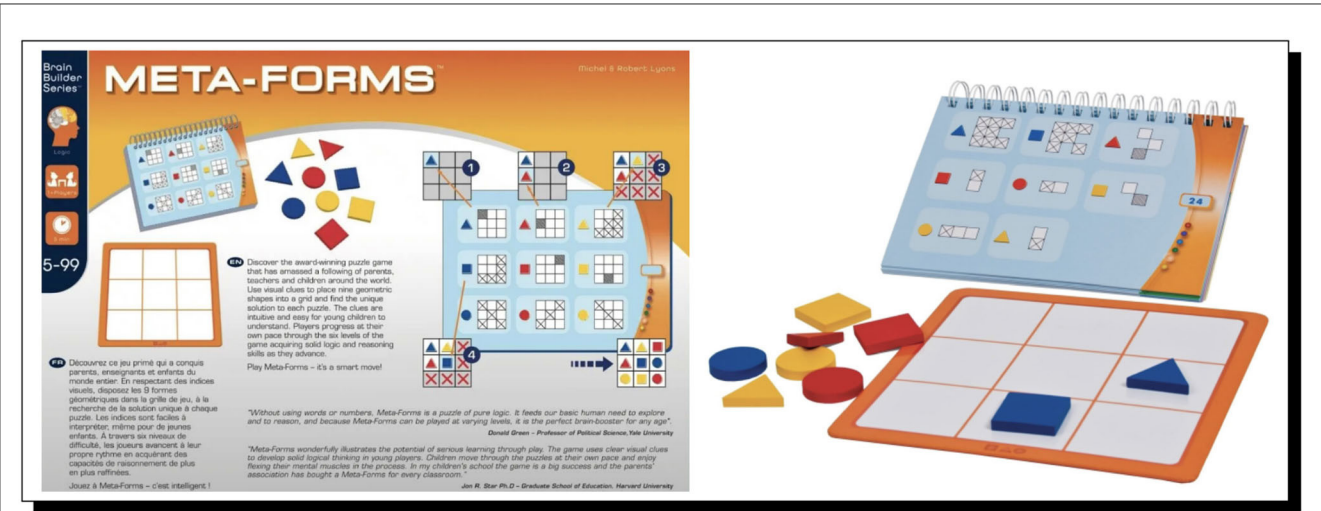


FIGURE 2
The Meta-Forms game, from FoxMind. This game provides a series of “clues” to the would-be puzzle solver, each of which is a visual version of a “logical statement,” which, in turn for our AI work, becomes a formula in a cognitive calculus (often requiring for such logicization only the formal language of a standard extensional logic such as \mathcal{L}_1). The goal is to physically construct a complete configuration of the 3x3 board from these clues, i.e., a full placement of each of the nine different objects in the game (3D versions of a triangle, square, and circle, each of which can be one of the three colors of red, blue, and yellow). Formally, if Π is a complete configuration of the board, and Γ the collection of formulae that logicize all clues, necessarily $\Pi \cup \Gamma$ is provably consistent in \mathcal{L}_1 and more expressive logics that subsume it.

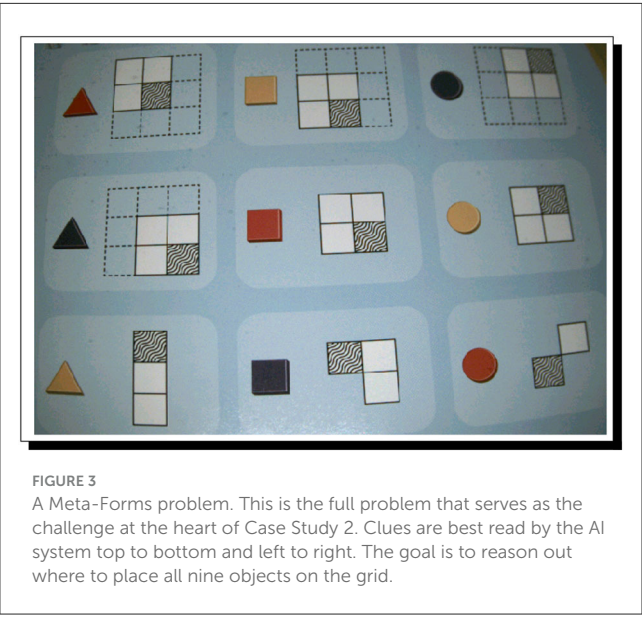


FIGURE 3
A Meta-Forms problem. This is the full problem that serves as the challenge at the heart of Case Study 2. Clues are best read by the AI system top to bottom and left to right. The goal is to reason out where to place all nine objects on the grid.

the case study, we give now some brief—but given purposes—additional relevant background on ARCADIA.

The ARCADIA cognitive architecture is composed primarily of a collection of non-introspectable processing units called *components*. On each ARCADIA processing cycle, components may take in and produce *interlingua content items*, which are tables of labeled data able to be interpreted by other components. Once generated, content items are placed in an accessible content area from which the architecture will select one on each processing cycle to become the *focus of attention*. This selected element is fed back into the components and used to generate more content

items. The strategy for selecting a content item is decided on a task-to-task basis that favors items, representing things deserving of more attention, such as those representing changes to objects within the field of vision. Though this architectural design and various types of components are motivated, as we have said, by the cognitive science of cognition, ARCADIA is able to smoothly and efficiently perform a robust range of tasks as implemented computation—such as object recognition, tracking, and driving (Bello and Bridewell, 2020).

To move into the case study, let us suppose that it is known that some people of interest are working on an unknown device in a building in an area that has a history of terrorist training and planning.³⁹ A team of “blue” artificial agents is tasked with deciding (and reporting to humans thereafter) whether or not these people of interest and the device with them pose a threat. The investigating team operates under the two-part assumption that those in the building are possibly terrorists, and the device in question possibly a bomb. In total, there are four investigative artificial agents. Three of them are in the vicinity of the building and are approaching it to ascertain the nature of the device in question via their sensors. These three agents are a high-altitude drone with a scanner (denoted by constant *hdrone*), a low-altitude drone with a camera (denoted by constant *ldrone*), and a land-based agent with wall-penetrating radar (denoted by constant *radar*). The final agent is a special argument-adjudicating agent (*adjudicator*) in full command of both cognitive calculi *DCEC* and *IDCEC* and also ShadowProver and ShadowAdjudicator; this agent is tasked with sending mission commands and receiving

39 This general premise is unfortunately far from implausible and is used as well in a simpler ARCADIA-less/perception-less adjudication scenario presented in Bringsjord et al. (2021), which is directly inspired by real events in the past.



FIGURE 4

PERI.2 observes the clue (left) and holds a Meta-Form piece in One Hand (center), correctly placing the shape (right). The clue, when logicized by PERI.2, can be represented as: $B[peri2, now, LocatedAt(bluesquare, 1) \vee LocatedAt(bluesquare, 2) \vee LocatedAt(bluesquare, 4) \vee LocatedAt(bluesquare, 5)]$. Notably, this is a disjunction. The challenge is to dynamically adjust arguments through time as clues are perceived by trying to negate disjuncts. Machine-vision middleware for PERI.2 is courtesy of Cognex, three of whose cameras are part of PERI.2 as well; hands are from Barrett Technologies.

```
(base) root@97e884a1add6:/base# python demos/2022_PERI2/peri_meta_forms.py
PERI.2 Reasoning About Blue Square's Location From Clues...
Proof found in 0.8197684288024902 seconds.
(class com.naveensundarg.shadow.prover.representations.formula.Belief
  [(:FOLFromSnark
    Givens:
    ((forall (?x ?y ?l) (implies (and (LocatedAt ?x ?l) (LocatedAt ?y ?l) ) (= ?x ?y)))

    ((LocatedAt yellowtriangle 1)
      (GIVEN[]))
    (not (= bluesquare yellowcircle))

    ((LocatedAt yellowcircle 2)
      (GIVEN[]))

    ((LocatedAt redtriangle 5)
      (GIVEN[]))
    (or (LocatedAt bluesquare 1) (LocatedAt bluesquare 2) (LocatedAt bluesquare 4) (LocatedAt bluesquare 5) )
    (not (= bluesquare redtriangle))
    (not (= bluesquare yellowtriangle)))
    Goals:
    (LocatedAt bluesquare 4)))
PERI.2 Has Justified Belief. Proving Truth...
Proof found in 0.2632763385772705 seconds.
(:FOLFromSnark
  Givens:
  ((AnswerKeyContains (LocatedAt bluesquare 4))
    (GIVEN[]))
  (forall (?x ?y) (implies (AnswerKeyContains (LocatedAt ?x ?y)) (LocatedAt ?x ?y))))
  Goals:
  (AnswerKeyContains (LocatedAt bluesquare 4)))
Therefore, PERI.2 Knows (= Has Justified True Belief) That The Blue Square Belongs in Location 4:
(Knows! peri2 now (LocatedAt bluesquare 4))
```

FIGURE 5

PERI.2 comes to know by reasoning that the Blue Square is at location #4. A rather long run of automated reasoning eventuates in PERI.2's coming to know that the blue square is at location #4. The proof given here provides justification for PERI.2's belief. It is, in fact, true that the blue square belongs to location 4. Therefore, in accordance with the conception of knowledge as justified true belief, where both belief and knowledge are allowed to vary in strength [in order to surmount the famous problem of [Gettier \(1963\)](#), as explained in [Bringsjord et al. \(2020b\)](#)], PERI.2 knows the correct placement.

messages from the other agents. From these messages, it is to use all its information at each time step to determine by reasoning if the people and the device are a threat. The other agents do not have full cognitive power (i.e., most of the cognitive verbs captured by both *DCEC* and *IDCEC* cannot be instantiated by their processing; e.g., these agents do not have the epistemic

“power” of *believing* and *knowing*); rather, they are only *perceptive* and *communicative agents*, able to focus on commands and changes in their environment and report their percepts to the adjudicator agent. The adjudicator agent is, thus, able to reason about the state of the world using the full ensemble of our calculi and automated reasoners, but the subsidiary agents are restricted to proper parts

of the cognitive calculi in question. Both *DCEC* and *IDCEC* have in their formal languages both a perception operator *P* and a communication operator *S*, read as “says” (see again as needed Section 6.2.1); but the operators in this pair for belief, knowledge, intention, and action are not available to the subsidiary agents.

For implementation of this scenario, we use the Minigrid environment (Chevalier-Boisvert et al., 2018): a virtual grid world in which we can model our artificial agents with limited field-of-view and perceptual impedances. Our house is represented as a structure enclosed by walls that block visual sensors but allow use of wall-penetrating radar. There is an opening in the house; it represents a garage in which the individuals are working on the mysterious device. The individuals under investigation and the device being worked on are represented by special tiles, as are perceptual disturbances such as dust clouds. At a high level, the situation can be observed playing out in our environment, as shown in Figure 8. Our agents on the scene (i.e., *hdrone*, *ldrone*, and *radar*) use instances of ARCADIA, while the adjudicator agent (*adjudicator*), again, is built atop ShadowAdjudicator (Giancola et al., 2020), which now, courtesy of a tie-in with ARCADIA, has scientifically serious capacity for both perception and argument-based reasoning.⁴⁰

Time in our implementation is conceptualized as adjudication timesteps and ARCADIA steps. On each reasoner cycle, a predetermined movement command is issued to each of *hdrone*, *ldrone*, and *radar* by *adjudicator* and received by a transceiver component that creates an interlingua item based on this command. The attentional strategy prioritizes these command items; they, thus, become the focus of attention. The agent’s movement effector component receives this command item and executes it. In parallel to this, ARCADIA’s robust attentional-and-visual system monitors for changes from the visual sensor; this sensor creates items from objects in the field of view. In the event a fully represented object in memory is perceived and becomes the focus of attention, it will be passed to the transceiver component, which will, in turn, send a message containing the agent’s perception to the *adjudicator* agent, which adds the information about the agent’s perception to its knowledge-base. This information includes whether a threat was perceived or not. After receiving a new percept, *adjudicator* will reason over the known percepts and return a belief about the situation, in particular, the degree of belief regarding whether a threat exists. The overarching pipeline is shown in Figure 9. These degrees correspond to the levels introduced earlier in the present study (see again, if needed, Table 1).

The situation plays out as follows and is presented in Figure 10. First, *hdrone* is issued orders to scan the building in a fly-by. It perceives the building but does not perceive any objects beyond this. From these percepts (or lack thereof in this case), *adjudicator* cannot determine whether there is a threat at this time-step, derived as a *counterbalanced* (recall again Section 7.1 and Table 1) belief as



FIGURE 6

A Full Trio of clues are fogged over. Fog (courtesy of a fog machine) has appeared in the RAIR Lab, and the results are not good perception-wise.

to whether or not there is a threat. In other words, at this point *adjudicator* is agnostic.

Next, the low-altitude drone (*ldrone*), in possession of a camera, receives orders to make an approach. As it obeys, its attention is focused on the people of interest and the device, but before the internal representation of the object can be fully assembled . . . a dust storm is kicked up, and this blocks *ldrone*’s visual sensors, which, in turn, nullifies its ability to have its visual component form representations of individuals or the device. Instead, it directs its attentional focus at the dust cloud itself; this blocks its view. These percepts of the people and cloud are sent back to *adjudicator*, which, at this point (rationally), maintains a *counterbalanced*/agnostic epistemic attitude regarding a threat/no-threat (i.e., re. *p*).⁴¹

At this point, the aforementioned ground-based agent with wall-penetrating radar (*radar*) is deployed to the side of the building. Its attention is drawn to two men located around the suspicious device. The ground-based agent reports these percepts to the adjudicator agent; it, accordingly, believes that there is *more likely than not* a threat present.

We explain in some detail the reasoning at *t*₂ below. The adjudicator uses its *Domain_Knowledge*, which contains general rules for the situation, such as how to prioritize the beliefs of each agent and the definitions of negative and 0 belief in this context. When combined with the percepts reported by the ARCADIA Agents (*IDCEC_KB_at_t2*), ShadowAdjudicator is able to use *IDCEC* inference schemata to derive the current threat level. More formally, where this notation is simply “pretty printed” from underlying code, the situation is as follows:

⁴⁰ As the reader by now knows, *DCEC* has a perception operator (and a communication operator), but they are not in and of themselves connected to any genuine mechanization of attention and perception that is, in turn, based on the science of attention and perception in computational cognitive science. Connecting to ARCADIA changes this in one fell swoop.

⁴¹ This agnosticism is, in part, based on the initial percepts of the people of interest in the garage.

```
(base) root@97e884a1add6:/base# python demos/2022_PERI2/peri_meta_forms.py
Prover Output:
FAILED
```

FIGURE 7

PERI.2 fails to find a proof when perception is compromised. Due to fog in the environment, some key clues are now absent in automated reasoning, and there is failure because PERI.2 cannot turn disjunctive (indeterminate) clues into knowledge.

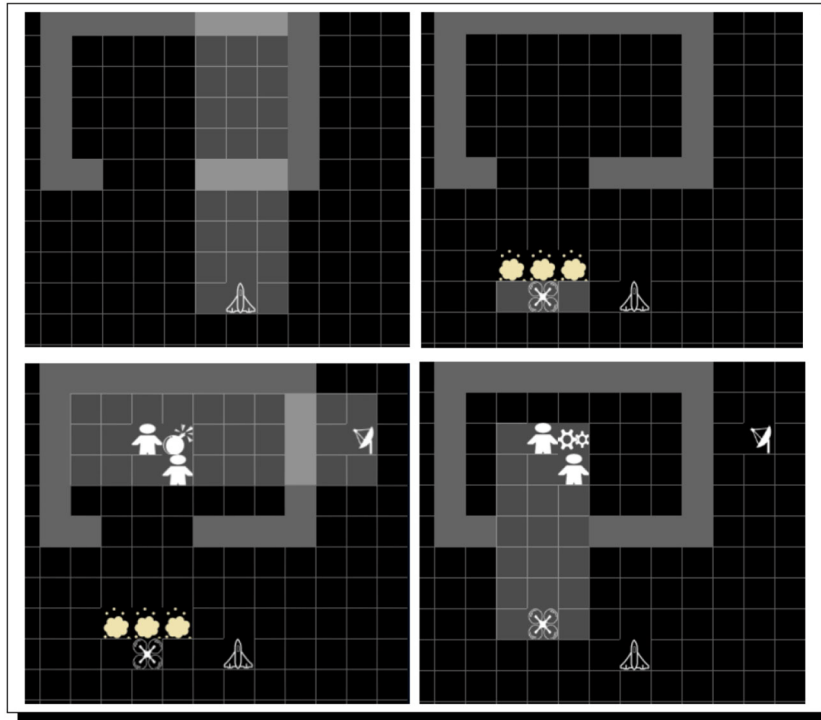


FIGURE 8

Multi-agent scanning. **(Top Left)** The high-altitude drone scans the building but does not perceive anything. **(Top Right)** The low-altitude drone moves in, but before attention can be focused on the objects in the building, a dust storm blocks its visual sensors. **(Bottom Left)** A ground-based agent with ground-penetrating radar moves into position and scans the inside of the building. **(Bottom Right)** The dust cloud disappears and the low-altitude drone's attention is drawn to the open building, where it perceives two men benignly working on an engine.

$$\begin{aligned} \text{Domain_Knowledge} = & \{\forall t_0, t_1, t_2 : \mathbf{B}^h(\text{hdrone}, t_0, \phi) \\ & \wedge \mathbf{B}^r(\text{radar}, t_1, \phi) \wedge \mathbf{B}^l(\text{ldrone}, t_2, \phi) \Rightarrow \\ & \mathbf{B}^{\max(r-1/4, l-1/4, l-1/2)}(\text{adjudicator}, \max(t_0, t_1, t_2), \phi), \\ & \forall t : \mathbf{B}^\sigma(\text{adjudicator}, t, \neg\phi)\} \\ \Leftrightarrow & \mathbf{B}^{-\sigma}(\text{adjudicator}, t, \phi), \\ & \forall t : \forall a : \neg\mathbf{P}(a, t, \neg\phi) \wedge \neg\mathbf{P}(a, t, \phi) \Rightarrow \mathbf{B}^0(a, t, \phi)\} \\ \text{IDCEC_KB_at_t2} = & \{\neg\mathbf{P}(\text{hdrone}, t_0, \neg p), \neg\mathbf{P}(\text{hdrone}, t_0, p), \\ & \neg\mathbf{P}(\text{ldrone}, t_1, \neg p), \neg\mathbf{P}(\text{ldrone}, t_1, p) \\ & \mathbf{P}(\text{radar}, t_2, \neg p)\} \\ \text{Domain_Knowledge} \cup \text{IDCEC_KB_at_t2} \vdash_{\text{IDCEC}} & \mathbf{B}^1(\text{adjudicator}, t_2, p) \end{aligned}$$

Finally, the low-altitude drone (*ldrone*) manages to emerge from the dust storm after new orders and is thus once again able to observe into the building. It focuses its attention on the device and

... perceives it to be a benign car engine. Once this information is relayed back to *adjudicator*, it reasons that it is *unlikely* there is a threat.

It should be noted here that *adjudicator* has situation-dependent definitions within its knowledge-base and is able to perform perception-infused reasoning that factors in these formulae. For example, notably, the true percept reported to the adjudicator is not really the presence of threat proposition p as simplifyingly shown in $\mathbf{P}(\cdot, \cdot, p)$, as shown in Figure 10, but rather a percept of the true object that the agent perceives [in this case that of *hdrone*, $\mathbf{P}(\text{hdrone}, t_0, \text{wall})$]. From this, *adjudicator* uses domain-context knowledge with the given percept to determine whether the agent perceived a threat or if not enough was perceived to ascertain whether the agent perceived a threat or not. Additionally, this extends to the *adjudicator* having a context-aware understanding of different types of agents and different levels of perception power, some being stronger than others, which is why the visual sensor on *ldrone* overrides the

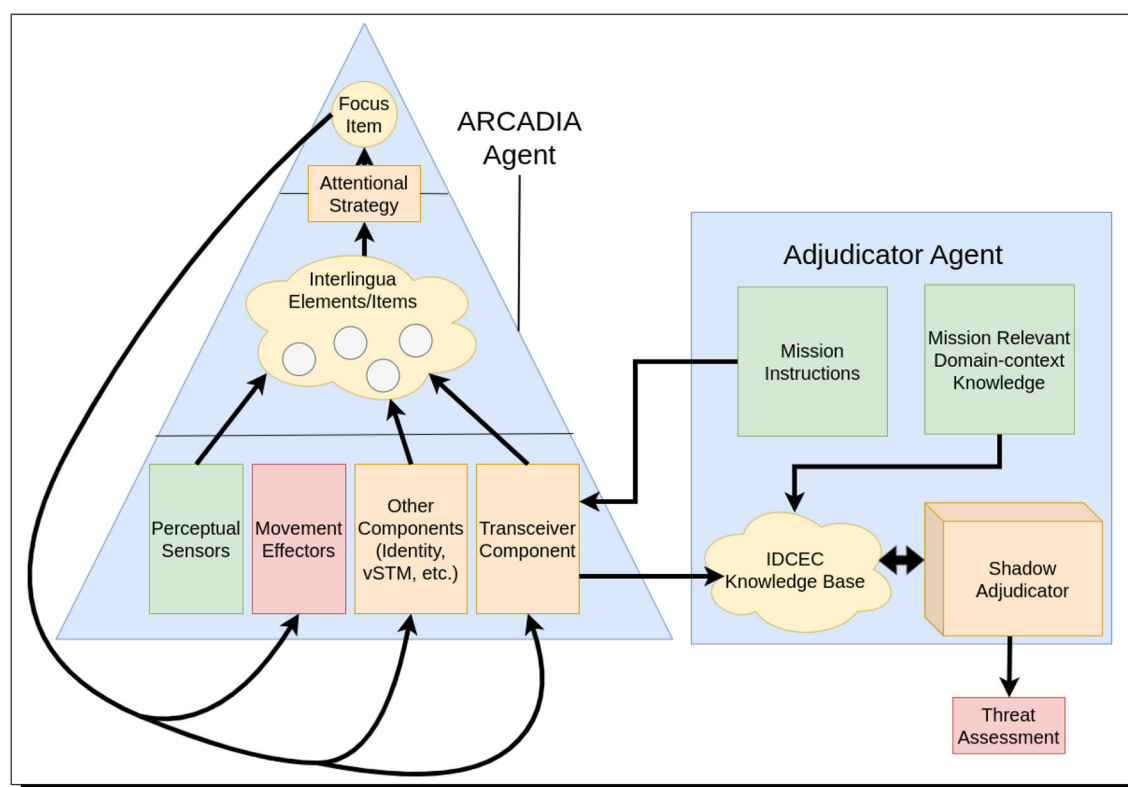


FIGURE 9

The information pipeline between the ARCADIA agent and the adjudicator agent is shown here. The high-level information pipeline between ARCADIA agents and the adjudicator agent is shown here. At each time step, mission instructions are passed to the ARCADIA agent in the situation via the agent's transceiver component. These commands are attended to and passed to the agent's movement effectors. The ARCADIA agent's perceptual sensors (visual, radar, etc.) pick out new items attended via the visual components that create objects. The finalized objects are interpreted to be fully perceived and are sent to the Adjudicator via the transceiver. The Adjudicator adjudicates between arguments factoring in the percepts of multiple agents on the ground, along with mission-relevant domain-context knowledge, to determine if there is a threat.

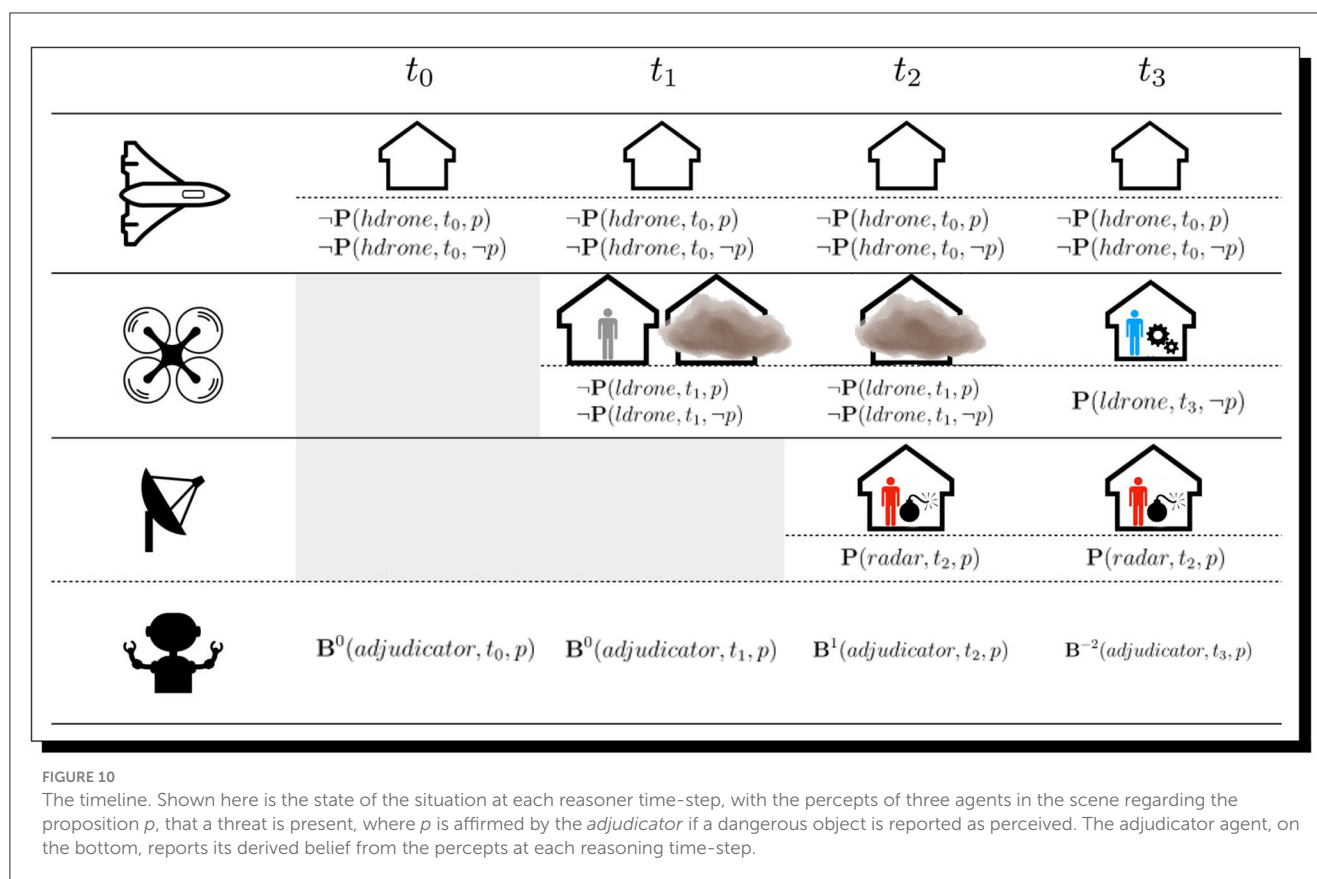
perceptions from *radar* at t_2 and t_3 . This event leads to its final belief at the *unlikely* level. This is also why the percepts from the wall-penetrating radar only lead to a *more likely than not* level of belief, rather than a belief at the level of *likely* at t_2 .

Summing up, our third case study provides not only a potential real-life example in which our automated argumentation systems play a central and salutary role, but also demonstrates that our system has many capabilities outlined in our desiderata, **Des**. In particular, Case Study 3 exemplifies the defeasible nature of our system as encapsulated by desiderata d_1 and the ability of our system to reason over cognitive operators as stated in desiderata d_6 . Regarding desiderata d_1 : As new information comes to light over the course of the scenario, the adjudicator is able to update its reasoning regarding the threat level at each time-step (see Figure 10); hence the reasoning capability of the system is observably defeasible, as desired. For desiderata d_6 , the system reasons over the cognitive operators for both belief and perception, as observed in depictions of both the agents on the scene and the adjudicator agent; see both Figure 10 and our presentation of Domain_Knowledge. This reasoning over cognitive operators also includes reasoning over the belief levels; hence part of desiderata d_3 is satisfied.

8 Sophistic argumentation

There is, it seems to us, a long-standing bias or presumption within the logicist AI tradition (into which, as explained above, our study as reported herein firmly falls) that treats arguments as fundamentally similar to earnestly constructed proofs (or at least to simplified, scaled-down proofs, earnestly and sincerely constructed). In this tradition, the purpose or function of arguments, like that of the authoring of proofs by humans engaged in the formal sciences, is to support rational belief fixation and to thereby enable new knowledge to arrive in the mind of cognizers who assimilate these proofs. This tradition makes room for and indeed realistically expects (at least periodically) invalid proofs (the history of mathematics having seen many), just as the tradition of computer programming makes plenty of room for programs that are invalid (but certainly programs).⁴² In point of fact, we ourselves, in adopting a thoroughgoing inference-theoretic perspective, regard arguments to be akin to proofs and

⁴² After all, program verification makes little sense if there cannot be programs that are invalid. For discussion of this, see Arkoudas and Bringsjord (2007). For a paradigm of program verification based directly on our brand of logicist AI, see Bringsjord (2015).



argument crafting on the part of humans to be akin to the craft of articulating proofs. However, while arguments do often function as demonstration and warrant in support of belief and decision-making, these are undeniably neither the sole functions of arguments nor are all warrants rational ones. This is something we suspect that AI should start to take note of, carefully. We, thus, now briefly explain, and our explanation will wrap up by drawing once again upon the three-door Monty Hall Problem = MHP₃, now familiar to our readers given earlier discussion of this problem.

To explain, let us first consider the function of arguments: Arguments are often instruments of persuasion. In fact, an argument's persuasiveness may be of greater import than its veracity or validity, depending on the arguer's intent with regard to its audience. Logician AI has largely followed in the footsteps of formal logic by privileging the dialectic (i.e., in a word, logic) over the other members of the ancient trivium. By eschewing rhetoric (essentially argumentation as treated today what is known as *informal logic*; see Groarke, 1996/2017), most logicist AI fails to appreciate the persuasive function of argument and its role in dialogical games such as disputation. This failure is not a small one. The persuasive power of argument is central to the practice of policy-making, politics, and law, and the life-altering decisions sometimes made therein. Moreover, persuasion is essential to the utility and success of logicist AI—even if this is unrecognized by practitioners. Why is it essential? Well, insofar as logicist AI in support of, and interacting with, humans is concerned, the goal is both to “be correct” and to “be believed;” systems that are correct but not believed are useless. Furthermore, we charitably assume

that acceptance and use of these logicist-AI systems are intended to be volitional, and as such, the goal again is to “be believed,” not simply to “be obeyed;” systems that are obeyed even when not believed are undesirable, dangerous, and potentially unethical.

Second, regarding rationality, arguments can be persuasive even when they are invalid or untruthful, and veracious arguments can be unpersuasive (as the literature on MHP₃ confirms; see the discussion of this empirical fact in Chapter 1 of Pinker, 2021). In terms of bringing about human belief, validity and veracity guarantee nothing. That invalid, pseudo-rational arguments can be persuasive is not a new revelation; Aristotle knew this over two millennia back when he wrote that arguments can have the appearance (but not always the substance) of demonstrable justification that makes belief warranted (Aristotle, 1823). Indeed, the methodological and disciplinary distinction between rhetoric and dialectic—between persuasion and veracity—dates back firmly and in general to ancient Greece and the age and work of not just Aristotle, but Plato and Socrates (see, specifically, the claimed intellectual battles between Socrates and the sophists).

Who were the sophists? To brutally summarize some of Plato's dialogues, the sophists were itinerant teachers who, for money, taught the skill of persuasive argumentation and debate to Athenian citizens so that they might prevail in the courts and in civic life—even if they were in the wrong. The sophists were criticized and opposed by Socrates and others because they (supposedly) only cared about being persuasive. They have been characterized as purveyors of the semblance of wisdom and not the genuine article, having rejected the doctrinal ideal of “truth” to promulgate, instead

of the virtue of persuasive cleverness without moral good (Aristotle, 1955). While many contemporary scholars (see Marback, 1999; Gagarin, 2001; McComiskey, 2002) have attempted to rehabilitate the sophists' reputation, the legacy of the sophists—among both scholars the general public—still amounts to “sophistry” being a byword for insincerity, self-interest, and, above all, manipulative persuasion by clever argumentation.

This encapsulated history of the sophists is given by us here for more than just trivia; the sophists demonstrated the power and importance of persuasion (viz. rhetoric), attempted to systematize it, and stand as a cautionary warning about the pursuit of argument-based persuasion unchecked by truth or virtue. However, why, the reader might ask, did the sophists' techniques work? More importantly, why are invalid arguments sometimes so persuasive? The answer to that is rather simple: Absent sufficient training and in-the-moment mental effort, humans are abysmal at normative argumentation and rational judgment. Humans are, unknowingly, imperfect reasoners who predictably and instinctively succumb to a host of biases and illusions and, moreover, are supremely, yet undeservedly, overconfident of their ability to reason and judge the reasoning of others—at least when compared with the standards of formal deductive and inductive logics and probability theory.

Moreover, the takeaway is that not only do logicist-AI systems need argumentation but also they need persuasive argumentation that ensures and preserves truthfulness (veracity) and formal validity in order to engender rational human use. Perhaps the reader will agree that we do not want artificial agents able to understand and generate arguments wonderfully, in order to, in part, persuade humans sophistically.

Before moving on to the final section of the present study, it is, in our view, worthwhile to say a bit more about the sophists, and to then end this section by looking at a specimen of just the sort of sophistic argumentation that AI systems should not produce and promote in order to persuade humans.

Naïve and unfair as their remembrance may be—the truth is that ancient sophistic techniques have been vibrantly alive and well and continuously refined for over two millennia—persuasive techniques that prey upon the audience's cognitive dissonance, ignorance, intellectual laziness, and desire for comforting belief reinforcement. Is there the specter of digital sophists emerging? Why yes. Sophistic AI is literally a past accomplishment. Starting in the early 2000s, the application of AI to natural argumentation refocused on audience-centric systems that take subjective aspects of argumentation seriously (see Reed and Grasso, 2001, 2007; Reed and Norman, 2004) and this resulted in the development of various neo-rhetorical (e.g., Grasso, 2002) and logico-dialectical (e.g., Aubry and Risch, 2006) approaches to persuasive and deceptive argumentation. In 2010, cognitive models were added to the mix, resulting in *The Lying Machine* (Clark, 2010), an explicitly sophistic artificial agent that persuades via a combination of argumentation and illusion.

The Lying Machine (TLM) is a logicist-AI system that manipulates human beliefs through persuasive argument by using cognitive models to generate convincing yet potentially disingenuous arguments. In design, the machine maintains conceptually separate repositories for its first- and second-order beliefs (i.e., its beliefs about the world and its beliefs about its

audience's beliefs about the world). It reasons over first-order beliefs in a normatively correct fashion, but when reasoning over second-order beliefs, it uses both normatively correct reasoning and a predictive theory of human reasoning, namely, *mental models* theory (Johnson-Laird, 1983, 2006), one of the most influential theories of human reasoning in cognitive science. In so doing, the machine internally contrasts (i) what it believes, (ii) what it believes its audience ought to believe were they to reason correctly, and (iii) what it believes its audience will likely believe given their predicted fallibility. In operation, TLM seeks to achieve various persuasion goals of the form “persuade the audience of ϕ ,” where ϕ is a logicization of a proposition $\langle\phi\rangle$ about the world. Given such a goal, the machine first forms its own justified belief about ϕ .⁴³ TLM, then, determines whether its audience ought to believe $\langle\phi\rangle$ and whether $\langle\phi\rangle$ can be justified in convincing fashion based solely on second-order beliefs (i.e., beliefs it ascribes to its audience). If so, the machine, then, constructs and articulates a credible argument for ϕ , presented then as an argument for $\langle\phi\rangle$.⁴⁴ Like the sophists, TLM aims for *perceived* credibility as opposed to objective, logical, or epistemological credibility. While its arguments may be logically valid or invalid, the importance is that they *appear* valid to its audience. Argument credibility is enhanced by limiting the initial premises to what the audience is believed by TLM to already believe. Moreover, since the machine is not constrained by logical validity, it is able to produce all of the following types of arguments:

- a veracious argument for a true proposition emanating from shared beliefs;
- a valid argument for a false proposition emanating from one or more false premises that the audience erroneously believes already;
- a fallacious argument for a true proposition (an expedient fiction for the fraudulent conveyance of a truth); and
- a fallacious argument for a false proposition (the most opprobrious form being one that insidiously passes from true premises to a false conclusion).

With the above repertoire in hand, the lying machine attempts to take on the pejorative mantle of the sophists by causing arbitrary belief to materialize in the minds of those targeted, through persuasive argumentation without concern for validity, sincerity, or truth. The results of experiments with TLM are, perhaps, unfortunate but not surprising, given that the fully replicated and thoroughly confirmed empirical fact of the matter in the cognitive science of reasoning has disclosed that humans confidently believe any number of things on the strength of reason that is often downright absurd, logically and mathematically speaking. [An excellent, if depressing, survey of this science is given in the study

⁴³ That is to say, it determines and internally justifies whether ϕ follows from, or is contradicted by, first-order beliefs (i.e., its own beliefs about the world), as regulated by background inference schemata (which obviously include normatively invalid ones, e.g., affirming the consequent).

⁴⁴ Natural-language-generation aspects of TLM are left aside here since out of scope.

by [Pinker \(2021\)](#), the anchoring first chapter of which features the very same MHP₃ problem first introduced in the present essay in Section 2.] Humans find the machine's sophistic arguments both credible and persuasive, even when those arguments are opposed by (logically) valid rebuttals ([Clark, 2010, 2011](#)).

We now end the present with an informal presentation of an argument regarding MHP₃ that practitioners of human-centric AI need to ensure is not generated, nor accepted, by artificial agents. The argument in question is in support of a policy of STAY in the problem, and runs as follows:

The Lame-Horse Argument

- (1) Suppose you bet at random on Horse #2 in a three-horse race, where all three horses at the outset are indistinguishable with respect to all of their respective racing-relevant properties.

(Of course, the idea is that in MHP₃ we have a three-door "race," and the bet is the initial selection of one of the three doors.)

- (2) From (1), we deduce that your odds of winning at t , the moment the race starts, are $\frac{1}{3}$.
- (3) Suppose as well that during the race, at $t'(t' > t)$, Horse #3 suddenly comes up lame and is out for good, while Horse #1 and Horse #2 continue running, neck and neck.
- (4) From (3), we deduce that your odds of winning at $t''(t'' > t')$, the moment after Horse #3 drops out, are $\frac{1}{2}$.
- (5) We can also infer that switching your bet to Horse #1 at the next instant $t'''(t''' > t'')$, with all conditions remaining the same (& assuming that you are given the opportunity to switch) is irrational, because the effort of doing so will not improve your $\frac{1}{2}$ odds at all.
- (6) Since the scenario here is isomorphic to that seen in MHP₃ (where of course your opportunity to switch doors is just like your opportunity to switch horses), it's irrational for you, or for that matter any contestant, to switch doors after Monty Hall reveals a donkey (or llama, etc.), a move that is of course the analog for Horse #3 coming up lame and thus "revealing" itself to be a guaranteed loser.

The Lame Horse Argument is a powerful sophistic argument; as [Pinker \(2021\)](#) explains, it even persuaded many professional mathematicians that a STAY policy in MHP₃ is irrational (an extensive treatment of, and references for, The Lame-Horse Argument, can be found in the study by [Granberg, 2014](#)). Of course, this is not to say that such mathematicians *intended* to persuade their targets while knowing that their argument was invalid. However, regardless, this is certainly something that could be done by malevolent agents (whether human or artificial), rather easily. Thus, if we may be so bold, the argument here is one that by our lights, the sophists would be quite happy with, in general; it is an argument, if you will, right up their alley.

However, *why* is The Lame-Horse Argument unsound? Though it is persuasive, it is not veracious because (in short), in point of fact, the two scenarios are not isomorphic at all (and that they

are is a premise in the argument); they are not even analogous by the simplest inference schemata for analogical argumentation.⁴⁵ The reason is that a number of intensional factors in the mind of Monty Hall himself are crucial to a correct, reasoned solution, but these factors are entirely absent from the three-horse scenario; these factors were discussed and logicized in the cognitive calculus *DCEC* in Section 7.1.⁴⁶

9 Next steps; conclusion

We now briefly describe a series of steps we are already in the process of taking, to further broaden and apply our approach. Readers both alert and knowledgeable will in the case of most if not all of them have already wondered whether our approach is applicable in these directions.

9.1 Surmounting the paradoxes of perception

The history of argument-based defeasible/non-monotonic systems in AI, as evidenced prominently by [Pollock \(1995\)](#), has been driven in no small part by the need to solve certain paradoxes, among which are the Lottery Paradox and the Paradox of the Preface.⁴⁷ Are there paradoxes specifically in the intersection of perception and such argumentation systems? Indeed there are; see for example the rather tricky one presented in [Davis \(1989\)](#). We are working hard on proving, and empirically demonstrating via simulations, that this and other even-harder paradoxes can be surmounted by our cognitive calculi and associated automated reasoners, in keeping with the desiderata that sum up our approach.

9.2 What about abductive argumentation?

Some of our readers will inevitably be curious about a type of reasoning we have yet to touch upon: *abductive* reasoning.⁴⁸ While

45 Laid out e.g., in [Bartha \(2013\)](#); [Bringsjord and Licato \(2015\)](#).

46 The three intensional prerequisites are: (i) Monty must *know* what's hidden by all doors; (ii) he must *perceive* and thereby come to *know* that initial choice; (iii) he must *intend* to open a losing door, and accordingly perform the associated action.

47 We do not fully agree with Pollock's proposed solutions to this pair of paradoxes, but such matters are out of scope presently.

48 Because (a) we momentarily provide information regarding how our approach will be extended into abductive reasoning, and (b) this information could not have been assimilated by the reader in advance of our laying out our approach, and instantiating it in the three case studies, we judged the present, concluding section to be the optimal location for our discussion of abduction. Notably, there are forms of abduction that in fact are not viewed as reasoning. This is nicely discussed in the study by [Douven \(2021\)](#), which begins with a key distinction: abduction viewed as the generation of hypotheses vs. abduction as the reasoning that justifies propositions, especially propositions that are hypotheses. Clearly, it is the latter form that is our concern.

it is certainly the case that there is no consensus as to what the precise nature of this reasoning is, the agreed-upon kernel of such reasoning in formal logic and AI expressed as an inference schema at least roughly in the fashion, followed earlier in the study, is as follows (where “ ϕ ” and “ ψ ” are formulae in accordance with some formal language, “ v ” denotes one or more variables free in these formulae, and χ denotes one or more constants/names):

$$\frac{\psi(\chi), \forall v[\phi(v) \rightarrow \psi(v)]}{\phi(\chi)}$$

Let us label this inference schema “ I_A .” This (deductively invalid, as desired) schema accords with many of the simple, familiar specimens of abduction. For example, suppose that soon after waking in the morning Bertram goes to the kitchen to make a cup of coffee, but upon entering the room finds a steaming cup of cappuccino sitting on his placemat at the breakfast table. No one else is present. Bertram asks himself: How did this situation come to be? Knowing that there is only one person—Abigail—in his household fully capable of making the exact kind of coffee he prefers, with knowledge of where he customarily sits, Bertram abduces via I_A , instantiated, to produce the following argument, to which Bertram accedes, and the mystery is solved (and he has gained knowledge as to whence the coffee cup).⁴⁹

The Abductive Coffee-Mystery Argument

1. $OnTable(cup22)$
2. $Prepared(abigail, cup22) \rightarrow OnTable(cup22)$
- \therefore 3. $Prepared(abigail, cup22)$

Unfortunately, as has been long and widely appreciated, I_A , and indeed any schema that is of this general sort, is deeply problematic. The set of defects has little to do with the mere (and desired) fact that abductive reasoning is non-deductive (it is, in this regard, a specific type of reasoning falling with inductive logic as the subdiscipline of logic our work falls into and is hence analyzed in the study by Johnson, 2016). For instance, this set of defects includes the havoc that can ensue from multiple uses of I_A : Let the universally quantified formula be instantiated twice (separately) to yield

$$\forall x[R(x) \rightarrow S(x)]$$

and

$$\forall x[\neg R(x) \rightarrow T(x)],$$

⁴⁹ Because abductive reasoning is often described as “inference to the best explanation”, and such inferencing is (plausibly, in our opinion) taken by many to be a cornerstone of the empirical sciences (see Douven, 2021), more elaborate examples from science could be given instead of our simple parable, but doing so is beyond scope and available space here—but we provide a few leads: For the reader not all that familiar with abduction, but with logic and science, in general, our recommendation is to read a seminal abductive model from Hintikka (1998). For those with an interest like ours, i.e., in human-centric AI and cognition, the place to start is without question the recently released Magnani (2023), and for a somewhat older but still-relevant overview of AI and computational logic, see Paul (2000).

and then suppose we have $S(a)$ and $T(a)$. A contradiction is, then, directly provable by two inferences, each in conformity with I_A .

Thus, one can view the chief challenge of working out a logic of abduction in the style of our cognitive calculi to be specifically the development of inference schemata that (i) are in the spirit of I_A , (ii) are (as it in fact is) machine checkable so that abductive argumentation is verifiable/falsifiable but (iii) have none of the obviously objectionable attributes of this inference schema. Of all the work we are aware of in this vein, Meheus and Batens (2006) comes closest to conforming to it and our approach. In this study, there is firm insistence upon having a proof theory, indeed one that is based on an attempt to expand and refine I_A . However, this proof theory could not be used to model and solve any of our three case studies. The reason is that the logic in question, LA^r , is purely extensional, as admitted by the researchers in question:

The logic presented in this study [LA^r] will be based on Classical Logic — henceforth CL. Moreover, all references to causality, laws of nature, and similar non-extensional concepts [such as belief, knowledge, and perception] will be out of the picture. We do not doubt that more interesting results may be obtained from intensional logics (Meheus and Batens, 2006, p. 22–223).

This quote can be viewed as a convenient stepping stone for a next step on our part, in which our cognitive calculi and automated reasoners, as introduced, explained, and deployed above, cover human-level abductive argumentation. The novel inference schemata in these calculi will minimally have perception and epistemic operators. Additionally, there would be a knowledge-base for the agent/s reasoning abductively. Thus, from our perspective, the coffee mystery is an enthymematic argument, both perceptually and epistemically. To achieve more precision, schema I_A would need to be expanded and refined; here, in fact, is a schema— I_A^{int} —marking a first such step in that direction, making use of the operators **B**, **K**, and **P** (for, as the reader will recall from the foregoing, belief, knowledge, and perception, respectively):

$$\frac{P(a, \psi(\chi)), K(a, \forall v[\phi(v) \rightarrow \psi(v)])}{B(a, \phi(\chi))}$$

This inference schema can formally and computationally undergird the argument Bertram might offer to someone as to why he regards the “mystery” to be solved, the idea being that he would express his reliance on *perceiving* the cup of cappuccino and his *knowing* beforehand the key conditional formula (and particular propositions re. Abigail), suitably instantiated. We are actively working on the expansion of our paradigm in this abductive direction.

9.3 What about pictorial argumentation?

Human agents make considerable use, even in sophisticated settings observed in the formal sciences, of arguments and proofs that include *pictorial* representations, where such representations are not reduced, and in some case not even in principle reducible to, symbolic content. [In our study described above (Case Study 2), we

have of course relied on the reduction of diagrams in Meta-Forms to linguistic formulae.] Notably, we are not here referring to arguments or proofs laid out in graphical ways (an important issue briefly discussed in Footnote 17). Reasoning frameworks, at least of the deductive sort that subsume extensional logics such as \mathcal{L}_1 and include both symbolic content (e.g., formulae in the formal language of a logic or—as in our case—cognitive calculi) and pictorial content, were seminally introduced by Barwise and Etchemendy (1995); they call such logics *heterogeneous*. Subsequently, a more general formal logic for heterogeneous reasoning, Vivid, was introduced by Arkoudas and Bringsjord (2009b). Vivid can be used to allow PERI.2 (and for that matter any logicist artificial agent) to reason about the Meta-Forms game board and clues relating to it as a diagram, unreduced to or represented by anything linguistic/symbolic. We are actively working on this direction, based on a new cognitive calculus with all the extant expressive and reasoning powers of \mathcal{DCEC} and \mathcal{IDCEC} and, at the same time, the vivid-like capacity to directly and irreducibly represent and allow reasoning over pictures, images, and diagrams.

9.4 Final words

We end by admitting that, at least in our view, the most daunting obstacle standing in the way of HCAI being based on argumentation science and engineering is not a technical one, at all. We are, for what it is worth, completely confident that the research trajectory explained (and hopefully rendered at least somewhat promising in the reader's view by virtue of the foregoing) above can indeed be used as the basis of artificial agents with near-human-level intelligence that profoundly help humans. However, humans have to *want* what argumentation-centric AI can provide. Our directive **Dir** is not (yet) universally affirmed. In a world where forms of AI, for instance large language models produced by so-called “Deep Learning,” wholly forego any argument or proof of the sort that we are calling for, we see room for plenty of rational concern. The forms we refer, as the reader will likely well-know, are purely statistical/connectionist ones entirely devoid of any declarative content expressed in accordance with a formal language (since they rely upon tokenization into formats that are only strings with none of the structure of quantification, inference schemata, etc.) and thus by definition devoid of any reasoning over such content in accordance with inference schemata.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

MC an expert on AI and sophistic argumentation and the automation thereof, principally contributed the vast majority of the section on this subject, and ensured that prose relating to this topic was suitably integrated across the essay. MC made contributions to many additional facets of the paper as well. PB chief architect and

developer of ARCADIA, oversaw and coordinated all integration between attention and perception on the one hand (as modeled in ARCADIA), and automated reasoning and automated planning on the other. He made contributions to many other facets of the paper as well. JO is one of the designers and principal writer of the ARCADIA-based case study, and leveraged his understanding of both ARCADIA and the RAIR Lab's automated reasoning/planning assets as well to engineer and describe the MiniGrid case study. JO also made contributions to many other facets of the paper. JS is the cognitive roboticist on the team; PERI.2 is overseen, engineered, and managed by him and his team in the RAIR Lab. Accordingly, the robot case study was enabled by JS and his efforts. JS made contributions to many other facets of the paper as well. NG is the long-time principal architect and developer of automated-reasoning and automated-planning systems in and from the RAIR Lab; in the case at hand, he originated the designs and code for both ShadowProver and Spectra (the former being part of the foundation of ShadowAdjudicator). NG is also the inventor of a number of cognitive calculi referred to in the paper. Overall, NG's work enables and infuses nearly all facets of the paper. MG with SB, was the principal writer of the paper, is the lead developer of ShadowAdjudicator, wrote with SB the “manifesto” part of the paper and propagated it throughout the essay, and used his expertise on MHPk to provide crucial content throughout the paper. MG worked directly on nearly every part of the paper, start to finish, engineered runs of ShadowAdjudicator and ShadowReasoner, and archived and presented parts of these runs in the paper. SB is the inventor of the first cognitive calculi, and, at least in part, of every cognitive calculus since the first appeared early in the 21st century. He worked with NG on automated reasoning and planning to lay the foundation for the project here, before it started. SB is an expert in computational inductive logic and defeasible inductive reasoning, cognitive likelihood (which he originated), conceived and designed the paper, wrote early drafts of it, and continued to write/edit nearly all content in subsequent drafts, through to the final version.

Acknowledgments

Four reviewers provided sagacious, insightful reaction to an earlier version of the study, and the authors thank them (any remaining foibles are due to us). The authors are grateful to: AFOSR for a DURIP award to SB and NG that has brought PERI.2 to physical life (award #FA9550-21-1-0185), to ONR for sponsorship of R&D devoted to meta-cognitively perceptive artificial agents (award #N00014-22-1-2201) that has brought ARCADIA researchers (among whom is found PB), together with RAIR-Lab researchers, and to both AFOSR (award #FA9550-17-1-0191) and ONR for longstanding support of R&D in automated reasoning, planning, and logic-based learning.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

References

- Adams, E. (1998). *A Primer of Probability Logic*. Stanford, CA: CSLI.
- Aristotle (1823). "Rhetoric," in *A New Translation of Aristotle's Rhetoric*, ed J. Gillies (London: T. Cadell).
- Aristotle (1955). "On sophistical refutations," in *Aristotle: On Sophistical Refutations, On Coming-To-Be and Passing-Away, On the Cosmos*, eds E. S. Forster, and D. J. Furley (Cambridge, MA: Harvard University Press), 11–155.
- Arkoudas, K., and Bringsjord, S. (2007). Computers, justification, and mathematical knowledge. *Minds Mach.* 17, 185–202. doi: 10.1007/s11023-007-9063-5
- Arkoudas, K., and Bringsjord, S. (2009a). Propositional attitudes and causation. *Int. J. Softw. Informat.* 3, 47–65.
- Arkoudas, K., and Bringsjord, S. (2009b). Vivid: an AI framework for heterogeneous problem solving. *Artif. Intell.* 173, 1367–1405. doi: 10.1016/j.artint.2009.06.002
- Artemov, S. (2008). The logic of justification. *Rev. Symb. Logic* 1, 477–513. doi: 10.1017/S1755020308090060
- Artemov, S., and Fitting, M. (2020). "Justification logic," in *The Stanford Encyclopedia of Philosophy*, ed R. Zalta. Available online at: <https://plato.stanford.edu/entries/logic-justification>
- Ashcraft, M., and Radvansky, G. (2013). *Cognition*, Pearson, London, UK. This is the 6th edition.
- Aubry, G., and Risch, V. (2006). Managing Deceitful Arguments with X-Logics, in "18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2006), 13–15 November 2006, Washington, DC, USA", IEEE Press, pp. 216–219.
- Bartha, P. (2013). "Analogy and analogical reasoning," in *The Stanford Encyclopedia of Philosophy*, ed E. Zalta. Available online at: <http://plato.stanford.edu/archives/fall2013/entries/reasoning-analogy>
- Barwise, J., and Etchemendy, J. (1995). "Heterogeneous logic," in *Diagrammatic Reasoning: Cognitive and Computational Perspectives*, eds J. Glasgow, N. Narayanan, and B. Chandrasekaran (Cambridge, MA: MIT Press), 211–234.
- Bello, P. F., and Bridewell, W. (2020). Attention and consciousness in intentional action: steps toward rich artificial agency. *J. Artif. Intell. Conscious.* 7, 15–24. doi: 10.1142/S2705078520500022
- Bench-Capon, T., and Dunne, P. (2007). Argumentation in artificial intelligence. *Artif. Intell.* 171, 619–641. doi: 10.1016/j.artint.2007.05.001
- Bench-Capon, T., Dunne, P., and Leng, P. (1992). "A dialogue game for dialectical interaction with expert systems," in *Proceedings of the 12th Annual Conference on Expert Systems and Their Applications*, 105–113. Available online at: <https://cgi.csc.liv.ac.uk/tbc/publications/avignon92.pdf>
- Berge, C. (1989). *Hypergraphs: Combinatorics of Finite Sets*. Amsterdam: Elsevier.
- Bergmann, M., Moor, J., and Nelson, J. (2013). *The Logic Book*, McGraw Hill, New York, NY. This is the 6th edition.
- Boolos, G. (1987). A curious inference. *J. Philos. Logic* 16, 1–12. doi: 10.1007/BF00250612
- Boolos, G. S., Burgess, J. P., and Jeffrey, R. C. (2003). *Computability and Logic*, 4th Edn. Cambridge: Cambridge University Press.
- Bretto, A. (2013). *Hypergraph Theory: An Introduction*. Cham: Springer.
- Brewka, G., Eiter, T., and Truszczynski, M. (2011). Answer set programming at a glance. *Commun. ACM* 54, 92–103. doi: 10.1145/2043174.2043195
- Bridewell, W., and Bello, P. F. (2015). "Incremental object perception in an attention-driven cognitive architecture," in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (Pasadena, CA: Cognitive Science Society), 279–284.
- Bringsjord, S. (1985). Are there set-theoretic worlds? *Analysis* 45, 64. doi: 10.1093/analys/45.1.64
- Bringsjord, S. (2012). Belief in the singularity is logically brittle. *J. Conscious. Stud.* 19, 14–20.
- Bringsjord, S. (2015). A vindication of program verification. *Hist. Philos. Logic* 36, 262–277. doi: 10.1080/01445340.2015.1065461
- Bringsjord, S., Giancola, M., and Govindarajulu, N. (2020a). "Culturally aware social robots that carry humans inside them, protected by defeasible argumentation systems," in *Culturally Sustainable Social Robotics, Proceedings of Robophilosophy 2020 August 18-21 2020*, eds M. Nørskov, J. Seibt, and Q. S. Quick (Amsterdam: IOS Press), 440–456. Available online at: <http://kryten.mm.rpi.edu/CulAwareSocRobotsDefArgSysRP2020CoverandTOC.pdf>
- Bringsjord, S., Giancola, M., and Govindarajulu, N. S. (2023c). Logic-Based Modeling of Cognition, in R. Sun, ed., "The Cambridge Handbook on Computational Cognitive Sciences", Cambridge University Press, Cambridge, UK, pp. 173–209. The URL here goes to an uncorrected preprint. <http://kryten.mm.rpi.edu/SBringsjordetalL-BMC121521.pdf>
- Bringsjord, S., Govindarajulu, N., and Bringsjord, A. (2023a). *Three-Dimensional Hypergraphical Natural Deduction. Book of Abstracts. Logic Colloquium 2023, European Summer Meeting of the Association for Symbolic Logic (ASL)*. Available online at: <http://kryten.mm.rpi.edu/SBringsjordetalL-BMC121521.pdf>
- Bringsjord, S., Govindarajulu, N., and Giancola, M. (2021). automated argument adjudication to solve ethical problems in multi-agent environments. *Paladyn* 12, 310–335. doi: 10.1515/pjbr-2021-0009
- Bringsjord, S., and Govindarajulu, N. S. (2018). "Artificial intelligence," in *The Stanford Encyclopedia of Philosophy*, ed E. Zalta. Available online at: <https://plato.stanford.edu/entries/artificial-intelligence>
- Bringsjord, S., and Govindarajulu, N. S. (2020). Rectifying the mischaracterization of logic by mental model theorists. *Cogn. Sci.* 44, e12898. doi: 10.1111/cogs.12898
- Bringsjord, S., Govindarajulu, N. S., Licato, J., and Giancola, M. (2020b). "Learning Ex Nihilo," in *GCAI 2020. 6th Global Conference on Artificial Intelligence, volume 72 of EPIc Series in Computing*, International Conferences on Logic and Artificial Intelligence at Zhejiang University (ZJULogAI) (Manchester: EasyChair Ltd.), 1–27. Available online at: <https://easychair.org/publications/paper/NzWG>
- Bringsjord, S., Govindarajulu, N. S., and Oswald, J. (2023b). "Universal cognitive intelligence, from cognitive consciousness, and lambda (\wedge)," in "Computational Approaches to Conscious Artificial Intelligence", Vol. 5 of *Machine Consciousness*, ed A. Chella (Singapore: World Scientific Publishing). Available online at: <http://kryten.mm.rpi.edu/ch5-main.pdf>
- Bringsjord, S., Govindarajulu, N. S., Slowik, J., Oswald, J., Giancola, M., Angel, J., et al. (2022a). "PERI.2 goes to preschool and beyond, in search of AGI," in *Proceedings of Artificial General Intelligence 2022 AGI-2022*, eds B. Goertzel, M. Iklé, A. Potapov, and D. Ponomaryov (Cham: Springer). Available online at: <http://kryten.mm.rpi.edu/ch5-main.pdf>
- Bringsjord, S., Govindarajulu, N. S., Taylor, J., and Bringsjord, A. (2022b). Logic: a modern approach: beginning deductive logic via HyperSlate™ and HyperGrader™. Motalen, Troy, NY. Available online at: <http://kryten.mm.rpi.edu/PERI2GoesToPreschoolAGI2022.pdf>
- Bringsjord, S., Hendler, J., Govindarajulu, N., Ghosh, R., and Giancola, M. (2022c). "The (uncomputable!) meaning of ethically charged natural language, for robots, and us, from hypergraphical inferential semantics," in *Trustworthy Artificial-Intelligent Systems, Vol. 102, textitIntelligent Systems, Control and Automation: Science and Engineering*, ed I. Ferreira (Cham: Springer), 143–167. Available online at: <http://www.logicamodernapproach.com>
- Bringsjord, S., and Licato, J. (2015). By Disanalogy, cyberwarfare is utterly new. *Philos. Technol.* 28, 339–358. doi: 10.1007/s13347-015-0194-y
- Bringsjord, S., Licato, J., Govindarajulu, N., Ghosh, R., and Sen, A. (2015). "Real robots that pass tests of self-consciousness," in *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015)* (New York, NY: IEEE), 498–504. Available online at: <http://kryten.mm.rpi.edu/SBringsjordetalsel-conrobotsg40601151615NY.pdf>
- Bringsjord, S., and Schimanski, B. (2003). "What is artificial intelligence? Psychometric AI as an answer," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)* (San Francisco, CA: Morgan Kaufmann), 887–893. Available online at: <http://kryten.mm.rpi.edu/scb.bs.pai.ijcai03.pdf>
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv*. Available online at: <https://arxiv.org/abs/2303.12712>
- Buss, S. (1994). On Gödel's theorems on lengths of proofs. I. Number of lines and speedup for arithmetics. *J. Symb. Logic* 59, 737–756. doi: 10.2307/2275906
- Buss, S. (1995). "On Gödel's theorems on lengths of proofs II: lower bounds for recognizing k -symbol provability," in *Feasible Mathematics II*, eds P. Clote, and J. Remmel (Basel: Birkhäuser), 57–90.
- Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

- Cerutti, F., Gaggli, S. A., Thimm, M., and Wallner, J. (2017). "Foundations of implementations for formal argumentation," in *The JfCoLog Journal of Logics and their Applications; Special Issue Formal Argumentation, Vol. 4*, eds P. Baroni, D. Gabbay, M. Giacomini, and L. Van der Torre (Cardiff: College Publications), 2623–2705.
- Chevalier-Boisvert, M., Willems, L., and Pal, S. (2018). *Minimalistic Gridworld Environment for Gymnasium*. Available online at: <https://github.com/Farama-Foundation/Minigrid>
- Clark, M. (2010). *Cognitive Illusions and the Lying Machine: A Blueprint for Sophistic Mendacity* (PhD thesis). Rensselaer Polytechnic Institute, Troy, NY, United States.
- Clark, M. (2011). "Mendacity and deception: uses and abuses of common ground," in *Building Representations of Common Ground with Intelligent Agents: Papers from the AAAI Fall Symposium*, eds S. Blisard, and W. Frost (Arlington, VA: AAAI Press). Technical Report FS-11-02, 2–9.
- Davis, E. (1989). *Solutions to a Paradox of Perception With Limited Acuity*. San Mateo, CA: Morgan Kaufmann Publishers, 79–82.
- Davis, E. (2017). Logical formalizations of commonsense reasoning: a survey. *J. Artif. Intell. Res.* 59, 651–723. doi: 10.1613/jair.5339
- Dietz, E., Kakas, A., and Michael, L. (2022). Argumentation: a calculus for human-centric AI. *Front. Artif. Intell.* 5, 955579. doi: 10.3389/frai.2022.955579
- Douven, I. (2011/2021). "Abduction," in *The Stanford Encyclopedia of Philosophy*, ed E. Zalta. Available online at: <https://plato.stanford.edu/entries/natural-deduction>
- Dung, P. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and N-person games. *Artif. Intell.* 77, 321–357. doi: 10.1016/0004-3702(94)00041-X
- Ebbinghaus, H. D., Flum, J., and Thomas, W. (1994). *Mathematical Logic*, 2nd Edn. New York, NY: Springer-Verlag.
- Fitting, M. (2015). "Intensional logic," in *The Stanford Encyclopedia of Philosophy*, ed E. Zalta. Available online at: <https://plato.stanford.edu/entries/logic-intensional>
- Francez, N. (2015). *Proof-Theoretic Semantics*. London: College Publications.
- Friedman, D. (1998). Monty Hall's three doors: construction and deconstruction of a choice anomaly. *Am. Econ. Rev.* 88, 933–946.
- Gagarin, M. (2001). Did the sophists aim to persuade? *Rhetorica*. 19, 275–291. doi: 10.1525/rh.2001.19.3.275
- Ganascia, J.-G. (2007). Modeling ethical rules of lying with answer set programming. *Ethics Inf. Technol.* 9, 39–47. doi: 10.1007/s10676-006-9134-y
- Gettier, E. (1963). Is justified true belief knowledge? *Analysis* 23, 121–123. doi: 10.1093/analys/23.6.121
- Giancola, M. (2023). *Reasoning with Likelihood for Artificially-Intelligent Agents: Formalization & Implementation* (PhD thesis). Troy, NY: Rensselaer Polytechnic Institute.
- Giancola, M., Bringsjord, S., Govindarajulu, N. S., and Varela, C. (2020). "Ethical reasoning for autonomous agents under uncertainty," in *Smart Living and Quality Health with Robots, Proceedings of ICRES 2020*, eds M. Tokhi, M. Ferreira, N. Govindarajulu, M. Silva, E. Kadar, J. Wang, et al. (London: CLAWAR), 26–41. Available online at: <https://github.com/RAIRLab/ShadowAdjudicator>; <http://kryten.mmm.rpi.edu/MGSBNSGCV/LogicizationMiracleOnHudson.pdf>
- Glymour, C. (1992). *Thinking Things Through*. Cambridge, MA: MIT Press.
- Gottwald, S. (2000/2015). "Many-valued logics," in *The Stanford Encyclopedia of Philosophy*, ed E. Zalta. Available online at: <https://plato.stanford.edu/entries/logic-manyvalued>
- Govindarajulu, N. S., Bringsjord, S., and Taylor, J. (2015). Proof verification and proof discovery for relativity. *Synthese* 192, 2077–2094. doi: 10.1007/s11229-014-0424-3
- Govindarajulu, N., and Bringsjord, S. (2017a). "On automating the doctrine of double effect," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, International Joint Conferences on Artificial Intelligence (Melbourne, VIC), ed C. Sierra, 4722–4730.
- Govindarajulu, N., Bringsjord, S., and Peveler, M. (2019). "On quantified modal theorem proving for modeling ethics," in *Proceedings of the Second International Workshop on Automated Reasoning: Challenges, Applications, Directions, Exemplary Achievements (ARCADE 2019)*, Volume 311 of *Electronic Proceedings in Theoretical Computer Science*, eds M. Suda, and S. Winkler (Waterloo, NSW: Open Publishing Association), 43–49. Available online at: <http://eptcs.web.cse.unsw.edu.au/paper.cgi?ARCADE2019.7.pdf>
- Govindarajulu, N. S., and Bringsjord, S. (2017b). "Strength factors: an uncertainty system for quantified modal logic," in *Proceedings of the IJCAI Workshop on "Logical Foundations for Uncertainty and Machine Learning" (LFU-2017)*, eds V. Belle, J. Cussens, M. Finger, L. Godo, H. Prade, and G. Qi (Melbourne, VIC), 34–40. Available online at: <http://homepages.inf.ed.ac.uk/vbelle/workshops/lfu17/proc.pdf>
- Granberg, D. (2014). *The Monty Hall Dilemma: A Cognitive Illusion Excellence*. Salt Lake City, UT: Lumad Press.
- Grasso, F. (2002). Would I Lie To You? Fairness and Deception in Rhetorical Dialogues, in R. Falcone & L. Korba, eds, "Working Notes of the AAMAS 2002 Workshop on Deception, Fraud and Trust in Agent Societies", Bologna, Italy. Held in conjunction with the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2002).
- Groarke, L. (1996/2017). "Informal logic," in *The Stanford Encyclopedia of Philosophy*, ed E. Zalta. Available online at: <https://plato.stanford.edu/entries/logic-informal>
- Heaven, W. D. (2022). Why meta's latest large language model survived only three days online. *MIT Technol. Rev.* Available online at: <https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/>
- Hintikka, J. (1998). What is abduction? The fundamental problem of contemporary epistemology. *Transact. Charles S. Peirce Soc.* 34, 503–533.
- Johnson, G. (2016). *Argument & Inference: An Introduction to Inductive Logic*. Cambridge, MA: MIT Press.
- Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. N. (2006). *How We Reason*. New York, NY: Oxford University Press.
- Kahneman, D. (2013). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus, and Giroux.
- Keynes, J. M. (1921). *A Treatise on Probability*. London: Macmillan.
- Kleene, S. (1938). On notation for ordinal numbers. *J. Symb. Log.* 3, 150–155. doi: 10.2307/2267778
- Kolmogorov, A. (1932). Zur Deutung der Intuitionistischen Logik. *Math Zeitschrift* 35, 58–65. doi: 10.1007/BF01186549
- Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitrechnung. Ergebnisse Der Mathematik. Translated as Foundations of Probability*. New York, NY: Chelsea Publishing Company.
- Koons, R. (2017). "Defeasible reasoning," in *The Stanford Encyclopedia of Philosophy*, ed E. Zalta. Available online at: <https://plato.stanford.edu/entries/reasoning-defeasible/index.html>
- Lenzen, W. (2004). "Leibniz's logic," in *Handbook of the History of Logic*, eds D. Gabbay, J. Woods, and A. Kanamori (Amsterdam: Elsevier), 1–83.
- Levesque, H. (2012). *Thinking as Computation*. Cambridge, MA: MIT Press.
- Levesque, H., and Brachman, R. (1985). "A fundamental tradeoff in knowledge representation and reasoning (revised version)," in *Readings in Knowledge Representation*, eds R. J. Brachman and H. J. Levesque (Los Altos, CA: Morgan Kaufmann), 41–70.
- Levesque, H., and Lakemeyer, G. (2007). "Chapter 24: cognitive robotics," in *Handbook of Knowledge Representation* (Amsterdam: Elsevier), 869–882. Available online at: <http://www.cs.toronto.edu/~hector/Papers/cogrob.pdf>
- Lorenzen, P. (1960). "Logic and agon," in *Atti del XII Congresso Internazionale di Filosofia IV* (Venice), 187–194.
- Luger, G. (2008). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, 6th Edn. London: Pearson.
- Łukasiewicz, J. (1920). O Logice Trojwartosciowej. *Ruch Filozoficzny* 5, 170–171.
- Magnani, L. (ed). (2023). *Handbook of Abductive Cognition*. Berlin: Springer Nature.
- Marback, R. (1999). *Plato's Dream of Sophistry*. Columbia, SC: University of South Carolina Press.
- McCarthy, J. (1980). Circumscription—a form of non-monotonic reasoning. *Artif. Intell.* 13, 27–39. doi: 10.21236/ADA086574
- McComiskey, B. (2002). *Gorgias and the New Sophistic Rhetoric*. Carbondale, IL: Southern Illinois University Press.
- McKeon, R. (ed). (1941). *The Basic Works of Aristotle*. New York, NY: Random House.
- Meheus, J., and Batens, D. (2006). A formal logic for abductive reasoning. *Logic J. IGPL* 14, 221–236. doi: 10.1093/jigpal/jzk015
- Modgil, S., and Prakken, H. (2014). The ASPIC⁺ framework for structured argumentation: a tutorial. *Argum. Comp.* 5, 31–62. doi: 10.1080/19462166.2013.869766
- Nelson, M. (2015). "Propositional attitude reports," in *The Stanford Encyclopedia of Philosophy*, ed E. Zalta. Available online at: <https://plato.stanford.edu/entries/prop-attitude-reports>
- Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.
- Paleo, B. W. (2016). "Leibniz's characteristic universalis and calculus ratiocinator today," in *Death and Anti-Death, Volume 14: Four Decades After Michael Polanyi, Three*

- Centuries After G. W. Leibniz', ed C. Tandy (Ann Arbor, MI: Ria University Press), 313–332.
- Paris, J., and Vencovská, A. (2015). *Pure Inductive Logic*. Cambridge: Cambridge University Press.
- Paul, G. (2000). "AI approaches to abduction," in *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 4*, eds D. Gabbay, and P. Smets (Dordrecht: Kluwer), 35–98.
- Pinker, S. (2021). *Rationality: What It Is, Why It Seems Scarce, Why It Matters*. New York, NY: Penguin Books.
- Pollock, J. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge, MA: MIT Press.
- Prakken, H., and Vreeswijk, G. (2001). "Logics for defeasible argumentation," in *Handbook of Philosophical Logic*, eds D. Gabbay, and F. Guenther (Dordrecht: Springer), 219–318.
- Prawitz, D. (1972). "The philosophical position of proof theory," in *Contemporary Philosophy in Scandinavia*, eds R. E. Olson, and A. M. Paul (Baltimore, MD: Johns Hopkins Press), 123–134.
- Reed, C., and Grasso, F. (2001). "Computational models of natural language argument," in 'Computational Science – ICCS 2001: International Conference San Francisco, CA, USA, May 28–30, 2001 Proceedings, Part I', Vol. 2073 of *Lecture Notes in Computer Science*, eds V. N. Alexandrov, J. J. Dongarra, B. A. Juliano, R. S. Renner, and C. J. K. Tan (Berlin/Heidelberg: Springer), 999–1008.
- Reed, C., and Grasso, F. (2007). Recent advances in computational models of natural argument. *Int. J. Intell. Syst.* 22, 1–15. doi: 10.1002/int.20187
- Reed, C., and Norman, T. J. (eds). (2004). *Argumentation Machines: New Frontiers in Argument and Computation*. Dordrecht: Kluwer Academic Publishers.
- Reiter, R. (1980). A logic for default reasoning. *Artif. Intell.* 13, 81–132. doi: 10.1016/0004-3702(80)90014-4
- Robinson, A. (1996). *Non-Standard Analysis*. Princeton, NJ: Princeton University Press.
- Rosenthal, J. (2008). Monty Hall, Monty Fall, Monty Crawl. *Math Horizons* 16, 5–7. doi: 10.1080/10724117.2008.11974778
- Russell, S., and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*, 3rd Edn. Upper Saddle River, NJ: Prentice Hall.
- Russell, S., and Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*, 4th Edn. Upper Saddle River, NJ: Prentice Hall.
- Seeger, K. (1967). Some modal logics based on a three-valued logic. *Theoria* 33, 53–71. doi: 10.1111/j.1755-2567.1967.tb00610.x
- Simon, H. (1956). Rational choice and the structure of the environment. *Psychol. Rev.* 63, 129–138. doi: 10.1037/h0042769
- Smith, R. (2017). "Aristotle's logic," in *The Stanford Encyclopedia of Philosophy*, ed E. Zalta. Available online at: <https://plato.stanford.edu/entries/aristotle-logic>
- Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., et al. (2021). Galactica: a large language model for science. *arXiv*. Available online at: <https://arxiv.org/abs/2211.09085>
- Tierney, J. (1991). *Behind Monty Hall's Doors: Puzzle, Debate and Answer?* New York, NY: The New York Times, 1.
- Toulmin, S. (2003). *The Uses of Argument*. Cambridge: Cambridge University Press.
- Walton, D., and Krabbe, E. (1995). *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. Albany, NY: State University of New York Press.

Frontiers in Artificial Intelligence

Explores the disruptive technological revolution of AI

A nexus for research in core and applied AI areas, this journal focuses on the enormous expansion of AI into aspects of modern life such as finance, law, medicine, agriculture, and human learning.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact

