Argumentation:
A Calculus for Human-Centric AI

Draft

May 26, 2022

Abstract

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.

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 [8] that could go beyond the ordinary human problem-solving capabilities within specific application domains, such as large-scale Planning [6], Data Analysis, and Data Mining.

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

[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 [30] describes the evolution of AI from the perspective of its link to human cognition, from its birth to today's developments.

ranging from organizing their daily routine, to ensuring compliance with legal or policy requirements, and 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, which include:

- "human in the loop" at the level of design, development, and deployment,
- human-friendly interaction within the sphere of Natural Language,
- cognitively 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 of HCAI systems in the human society,

and perhaps the most important desired characteristic of HCAI systems of

"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 in his paper "Computing Machinery and Intelligence" [59], a precursor of Artificial Intelligence:

[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`.

> "[...] 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 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 learning cannot typically give definitive results.

Importantly, 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, to facilitate the integration of the various processes and entities that exist within the application environment. The provision of such **cognitive explanations** forms the main requirement of today's **Explainable AI** and approaches such as Perspicuous Computing (`https://www.perspicuous-computing.science/`).

What is then an appropriate foundation for building HCAI systems with this variety of features and characteristics in terms of their operational behavior, and 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 challenges?

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 [16]. 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.

1.1 Argumentation in Cognitive Reasoning

Indeed, there is strong cognitive support for argumentation and its link to different cases of human thinking, stemming 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 [13]. 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 [21, 46] 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 [36], 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 [20] 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.

1.2 Argumentation as a Logical Foundation

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** [32, 50, 54] 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 [33]. These new logics aimed to capture the non-monotonicity feature of human reasoning, recognizing that, in contrast 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 [57, 47, 48], showed that it was possible to reformulate (and in some cases extend) most, if not all, such non-monotonic AI logical frameworks [5]. 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 [24, 22]. 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 Computation[4]), 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.

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

[4] `https://www.iospress.com/catalog/journals/argument-computation`

and use of HCAI systems that 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.

2 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 $\langle \mathcal{A}rgs, \mathcal{A}tt, \mathcal{D}ef \rangle$, where $\mathcal{A}rgs$ is a set of arguments, $\mathcal{A}tt$ is an **attack (or counter-argument)** binary relation between arguments, and $\mathcal{D}ef$ a **defense (or defeat)** binary relation between arguments. Typically, the defense relation $\mathcal{D}ef$ is a subset of the attack relation $\mathcal{A}tt$ capturing some notion of the relative strength between the attacking arguments. Hence when $(a_1, a_2) \in \mathcal{D}ef$ the argument a_1 is strong enough to defend

[5] Work in the area of Computational Argumentation can be found in the journal of Argument & Computation and the International Conference on Computational Models of Argument (COMMA). Other sources for review material of the area include [4, 55, 3, 62]

against (or defeat) a_2.

In practice, abstract frameworks are realized by structured argumentation frameworks [25, 15, 49, 43], 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** [65], 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 schemes in $\mathcal{A}s$ are parameterized named statements of association between different pieces of information. They can be represented in the simple form of $\mathcal{A}s = (\mathsf{Premises} \rhd \mathsf{Position})$, 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 coalition.[6] 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

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

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

2.1 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** [62, 66]. 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) [41]. 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 the contrastive element of the explanation. These arguments also point towards 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** [60] 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 [61] for a comprehensive review). These include studies of understanding the various types of argument schemes that humans use in their argumentative discourse [57, 65, 64], or how the process of human argumentation relates to human reasoning [48], 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 [60].

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 [52, 12]. 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 [17, 18]. 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 [29, 31] 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 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?

2.2 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 $\langle \mathcal{A}s, \mathcal{C}, \succ \rangle$? 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. 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:

$$\mathcal{A}s_1 = (HighScore \rhd Favorable) \qquad \mathcal{A}s_2 = (LowScore \rhd Non\_Favorable).$$

$HighScore$ 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 $Non\_Favorable$:

$$\mathcal{A}s_3 = (MiddleScore \rhd Favorable) \qquad \mathcal{A}s_4 = (MiddleScore \rhd Non\_Favorable).$$

To complete the representation of the problem, we include in the conflict relation the obvious conflict between arguments that support the incompatible positions $\mathit{Favorable}$ and $\mathit{Non\_Favorable}$, 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:

$$\mathcal{A}s_{+ve}(\mathit{Review}(\mathit{Id})) = (\mathit{Positive}(\mathit{Property}) \rhd \mathit{Favorable})$$
$$\mathcal{A}s_{-ve}(\mathit{Review}(\mathit{Id})) = (\mathit{Negative}(\mathit{Property}) \rhd \mathit{Non\_Favorable}).$$

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 $\mathcal{A}s_{-ve}(\mathit{Review}(\mathit{Id}))$ higher priority than (some of the) other arguments for the position $\mathit{Favorable}$. But, as the guideline indicates, this argument will not have higher priority than arguments built using the scheme $\mathcal{A}s_{+ve}(\mathit{Review}(\mathit{Id}))$ 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 a remains with a dilemma on being favorable or not towards 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 Consultant 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 [10] 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.

2.3 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 of their user group (e.g., junior specialized doctors in training, or general public for preliminary self-diagnosis).

Medical diagnostic knowledge that associates diseases with their observable symptoms can be represented in terms of argument schemes of the general form:

$$\mathcal{A}s = (\mathit{Symptoms} \rhd \mathit{Disease}).$$

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. Note that because these, expertly known, associations are treated as arguments, this 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:

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

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 [9] 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 diseases.[7] 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 towards the perspective of keeping the "human in the loop", where the systems aim to complement and strengthen the human expert's capabilities.

[7] 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.

3 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.

3.1 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 [42] aimed to capture the context sensitive nature of knowledge. Similarly, inheritance networks [19] 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 [53], aimed to capture the context-sensitive nature of commonsense reasoning with the knowledge of stereotypical sequences of events, and the change over time 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 [34].

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 [5].

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.

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 [62, 66, 23]. 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 [41]. 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 [11]? Several argumentation-based approaches study this question by considering how argumentative knowledge (arguments and strength) are extracted or mined from natural language repositories [29, 31], i.e., how argument schemes are formed out of text [65], or how we can recognize good quality arguments [17, 18] 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.

3.2 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 [33] 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 [39]) 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 [29, 31] 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 resolves 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 to 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, versus 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 [37, 38]. 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 [2, 58]. 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.

3.3 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 wired into the human mind.

For the classification of human experience and information processing mechanisms, Newell [45] 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" [45], 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 [1]. 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 (e.g., ACT-R [1], SOAR [27]) 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, Lebiere and Rosenboom [28] 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" [44] still seems to hold. Interestingly, this missing convergence towards unified theories of cognition persists across and within the *bands of cognition* [45]. 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 [28] emphasize that for human-like minds, the overall focus needs to be on 'the bounded rationality hypothesized to be central to human cognition' [56, **?**]. 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 towards this direction? 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. 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.

3.4 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 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 positions. 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 towards one position or stabilize to various (strong) positions that conflict with each other.

Another view on argumentation as a multi-agent system, following Mercier and Sperber [35], is to cast one agent as a communicator and another agent as the audience.[8] The exchange of information happens dynamically through the persuasiveness of the communicator and the *epistemic vigilance* of the audience.

In both cases, these environments need 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?

3.5 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.[9]

One such EU initiative is the publication of "Ethics Guidelines for Trustworthy AI"[10], 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 [51, 14]). 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, dis-

[8] This is also the basic mode of the method of dialectic argumentation in Aristotle, of a Questioner and an Answerer.

[9] 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`.

[10] `https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf`

pute, 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 [7] 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.[11] 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 3.1 and 3.2, 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 [63].[12]

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.[13] 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 towards "ethics by design", whereas the descriptive view would point towards an "evolutionary process". Adopting the

[11] https://www.moralmachine.net/

[12] Also consider `https://www.ai.rug.nl/~verheij/publications/oratie/oratie_Bart_Verheij.pdf`, `https://www.argnet.org/ethics-of-arg`

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

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 [39, 40].

Such a process of "ethics coaching", be it by the end 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 justifications 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 ones 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 these fundamental premises on which interlocutors disagree, even if it cannot help them reconcile their divergent views.

In his inaugural lecture, Verheij proposes not to regulate AI by enforcing human control or by the prohibition of 'killer robots', but through the use of argumentation systems which provide us with good arguments. It is not *the winning of a discussion but the finding of good answers to the difficult problems of life in a complex, dynamic world*, which forms intelligent behavior.[14]

4 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

[14] `https://www.ai.rug.nl/~verheij/publications/oratie/oratie_Bart_Verheij.pdf`

that offers an enhancement of peoples 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 towards 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" [26, 34]. 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. For the case of argumentation, we are fortunate to have already such a wide ranging study 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.

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 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?

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?

References

[1] John R. Anderson. *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press, 2007.

[2] Luis C.Lamb Artur S.d'Avila Garcez, Dov M.Gabbay. A neural cognitive model of argumentation with application to legal inference and decision making. *Journal of Applied Logic*, 12(2):109 – 127, 2014.

[3] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. Towards artificial argumentation. *AI Magazine*, 38(3):25–36, Oct. 2017.

[4] Trevor J. M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10-15):619–641, 2007.

[5] Andrei Bondarenko, Phan Minh Dung, Robert A. Kowalski, and Francesca Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artif. Intell.*, 93:63–101, 1997.

[6] Blai Bonet and Héctor Geffner. Planning as heuristic search. *Artificial Intelligence*, 129(1):5–33, 2001.

[7] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.

[8] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, Oxford, UK, 2014.

[9] Bruce G. Buchanan and Edward H. Shortliffe, editors. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, 1985.

[10] Oana Cocarascu and Francesca Toni. 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*, pages 9:1–9:8. ACM, 2016.

[11] Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15, 2009.

[12] Emmanuelle Dietz and Antonis C. Kakas. Cognitive argumentation and the selection task. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*, pages 1588–1594. Cognitive Science Society, 2021.

[13] Jonathan St B. T. Evans. *Thinking Twice: Two Minds in One Brain.* Oxford University Press, 2010.

[14] L. Floridi. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality.* OUP Oxford, 2014.

[15] Alejandro J. García and Guillermo R. Simari. Defeasible logic programming: an argumentative approach. *Theory and Practice of Logic Programming*, 4(1-2):95–138, 2004.

[16] Joseph Y. Halpern, Robert Harper, Moshe Y. Vardi, and Neil Immerman. On the unusual effectiveness of logic in computer science. *Bulletin of Symbolic Logic*, 7:1–19, 2001.

[17] Martin Hinton. Language and argument: a review of the field. *Research in Language*, 17(1):93–103, 2019.

[18] Martin Hinton. *Evaluating the Language of Argument.* Springer Nature Switzerland AG, 2021.

[19] John F. Horty, Richmond H. Thomason, and David S. Touretzky. A skeptical theory of inheritance in nonmonotonic semantic networks. *Artificial Intelligence*, 42(2):311–348, 1990.

[20] Daniel Kahneman. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York, 2011.

[21] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–91, 1979.

[22] A. C. Kakas. Informalizing formal logic. *Informal Logic*, 39(2):169–204, 2019.

[23] Antonis Kakas and Loizos Michael. Abduction and argumentation for explainable machine learning: A position survey, 2020.

[24] Antonis C. Kakas, Paolo Mancarella, and Francesca Toni. On argumentation logic and propositional logic. *Studia Logica*, 106(2):237–279, 2018.

[25] Antonis C. Kakas and Pavlos Moraitis. Argumentation based decision making for autonomous agents. In *Proc. of 2nd Int. Joint Conf. on Autonomous Agents & Multiagent Systems, AAMAS*, pages 883–890. ACM, 2003.

[26] Immanuel Kant. *Critique of Pure Reason.* The Cambridge Edition of the Works of Immanuel Kant. Cambridge University Press, New York, NY, 1998. Translated by Paul Guyer and Allen W. Wood.

[27] John E. Laird. *The Soar Cognitive Architecture.* The MIT Press, 2012.

[28] John E. Laird, Christian Lebiere, and Paul S. Rosenbloom. A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine*, 38(4):13, December 2017.

[29] John Lawrence and Chris Reed. Argument mining: A survey. *Comput. Linguistics*, 45(4):765–818, 2019.

[30] Antonio Lieto. *Cognitive Design for Artificial Minds.* Routledge, 2021.

[31] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10:1–10:25, 2016.

[32] Wiktor Marek and Mirosław Truszczyński. Autoepistemic logic. *J. ACM*, 38(3):587–618, 1991.

[33] John McCarthy. Programs with common sense. In *Semantic Information Processing*, pages 403–418. MIT Press, 1968.

[34] Drew McDermott. A critique of pure reason 1. *Computational Intelligence*, 3, 1990.

[35] Hugo Mercier and Dan Sperber. Intuitive and reflective inferences. *In Two Minds: Dual Processes and Beyond*, page 149–170, 2009.

[36] Hugo Mercier and Dan Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–74, 2011.

[37] Loizos Michael. Simultaneous Learning and Prediction. In *Proceedings of the 14th International Conference on the Principles of Knowledge Representation and Reasoning*, 2014.

[38] Loizos Michael. Cognitive Reasoning and Learning Mechanisms. In *Proceedings of the 4th International Workshop on Artificial Intelligence and Cognition*, pages 2–23, 2016.

[39] Loizos Michael. Machine Coaching. In *Proc. of IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)*, pages 80–86, 2019.

[40] Loizos Michael. Machine Ethics through Machine Coaching. In *Proc. of 2nd Workshop on Implementing Machine Ethics @ UCD*, 2020.

[41] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[42] Marvin Minsky. A framework for representing knowledge. In John Haugeland, editor, *Mind Design: Philosophy, Psychology, Artificial Intelligence*, pages 95–128. MIT Press, Cambridge, MA, 1981.

[43] Sanjay Modgil and Henry Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.

[44] A. Newell. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In *Visual information.* Academic Press, New York, 1973.

[45] Allen Newell. *Unified Theories of Cognition.* Harvard University Press, USA, 1990.

[46] Fabio Paglieri and Cristiano Castelfranchi. Why argue? towards a cost–benefit analysis of argumentation. *Argument & Computation*, 1(1):71–91, 2010.

[47] Ch. Perelman and L. Olbrechts-Tyteca. *The new rhetoric. A treatise on argumentation.* Notre Dame/ London: University of Notre Dame Press, 1969.

[48] J.L. Pollock. Defeasible reasoning. *Cognitive Science*, 11(4):481–518, 1987.

[49] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

[50] R. Reiter. A Logic for Default Reasoning. *Artificial Intelligence*, 13(1-2):81–132, 1980.

[51] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group, 2019.

[52] Emmanuelle-Anna Dietz Saldanha and Antonis C. Kakas. Cognitive argumentation for human syllogistic reasoning. *Künstliche Intell.*, 33(3):229–242, 2019.

[53] Roger C. Schank and Robert P. Abelson. Scripts, plans and knowledge. In PN Johnson-Laird and PC Wason, editors, *Thinking: Readings in Cognitive Science, Proceedings of the Fourth International Joint Conference on Artificial Intelligence*, pages 151–157. Tbilisi, USSR, 1975.

[54] Y. Shoham. Nonmonotonic logics: Meaning and utility. In *Proc. 10th International Joint Conf. on Artificial Intelligence (IJCAI-87)*, pages 388–393, Milan, 1987.

[55] Guillermo Ricardo Simari and Iyad Rahwan, editors. *Argumentation in Artificial Intelligence*. Springer, 2009.

[56] H. A. Simon. A behavioral model of rational choice. In *Models of Man, Social and Rational: Mathematical Essays on Ration- al Human Behavior in a Social Setting*. John Wiley and Sons., New York, 1957.

[57] S.E. Toumlin. *The Uses of Argument*. Cambridge University Press, 1958.

[58] Efthymia Tsamoura, Timothy Hospedales, and Loizos Michael. Neural-Symbolic Integration: A Compositional Perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5051–5060, 2021.

[59] A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[60] F. Van Eemeren, R. Grootendorst, and F.H. van Eemeren. *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*. Cambridge University Press, 2004.

[61] Frans H van Eemeren, Bart Garssen, Erik C. W Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M Wagemans. *Handbook of Argumentation Theory*. Springer Netherlands : Imprint: Springer, 1st ed. 2014. edition, 2014.

[62] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review*, 36, 2021.

[63] Bart Verheij. Formalizing value-guided argumentation for ethical systems design. *Artif. Intell. Law*, 24(4):387–407, 2016.

[64] D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.

[65] Douglas N. Walton. *Argumentation Schemes for Presumptive Reasoning*. Psychology Press, 1996.

[66] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative xai: A survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4392–4399. International Joint Conferences on Artificial Intelligence Organization, 2021.