Explainable Reasoning in Face of Contradictions: From Humans to Machines

Timotheus Kampik[1] and Dov Gabbay[2,3,4]

[1] Umeå University, Sweden
[2] University of Luxembourg, Luxembourg
[3] King's College London, United Kingdom
[4] Bar Ilan University, Israel
tkampik@cs.umu.se, dov.gabbay@kcl.ac.uk

Abstract. A well-studied trait of human reasoning and decision-making is the ability to not only make decisions in the presence of contradictions, but also to explain *why* a decision was made, in particular if a decision deviates from what is expected by an inquirer who requests the explanation. In this paper, we examine this phenomenon, which has been extensively explored by behavioral economics research, from the perspective of symbolic artificial intelligence. In particular, we introduce four levels of intelligent reasoning in face of contradictions, which we motivate from a microeconomics and behavioral economics perspective. We relate these principles to symbolic reasoning approaches, using abstract argumentation as an exemplary method. This allows us to ground the four levels in a body of related previous and ongoing research, which we use as a point of departure for outlining future research directions.

Keywords: Symbolic Artificial Intelligence · Explainable Artificial Intelligence · Non-monotonic Reasoning.

1 Introduction

Over the last decades, the public perception of what artificial intelligence is (and is not) has dramatically shifted. For example, in 1996 and 1997, when the reigning chess champion Gary Kasparov played against IBM's chess computer *Deep Blue*, the ability of playing chess well was considered a key characteristic of human intelligence. Today, as technically literate consumers can easily install a world champion-beating program on their mobile phones, the focus has shifted to other problems, which range from different games like Starcraft and Go to real-world challenges like fully autonomous driving in inner cities. Even the Turing test [31], which roughly speaking requires a machine to be able to deceive a human into thinking it is human, seems to fail the test of time; given current socio-technical information systems, distinguishing men from machines is increasingly challenging, even in contexts where the machine behavior is determined by simple scripts, for example when social media bots spread misinformation [28].

Hence, to define characteristics of intelligent behavior, more abstract approaches are required. Such approaches have, indeed, been introduced as principles of non-monotonic reasoning; most notably are relaxed forms of monotony, such as restricted

monotony [15] (also known as cautious monotony) and rational monotony [24][5]. From a symbolic artificial intelligence perspective, these properties are very useful because they can be formally verified. Still, the properties have some obvious limitations:

- The properties are merely *indicators* of intelligence; certainly, fairly "unintelligent agents" can also satisfy restricted monotony and rational monotony, simply by never inferring anything from any knowledge base.
- It is not clear how these properties relate to human intuitions of intelligence.

In this paper, we explore ways to address these limitations by *i)* building a conceptual bridge between formal principles of non-monotonic reasoning and empirical, as well as formal perspectives on human reasoning and decision-making and *ii)* illustrating how different formal approaches to non-monotonic reasoning reflect different levels of sophistication of human reasoning.

2 Human Intelligence: Bounded Rationality and Reasoning Backwards

As a preliminary for a bridge between human reasoning and formal methods of automated reasoning, let us provide a brief overview of the development of models of human reasoning and decision-making at the intersection of microeconomic theory and behavioral psychology. At least since the middle of the 20th century, studies in the fields of micro-economic theory and behavioral psychology attempt to identify patterns and *formal* models of human decision-making and reasoning, both for *descriptive* ("How do humans reason and make decisions?") and *prescriptive* ("How should humans reason and make decisions?") purposes[6]. An early theory that is still very influential is the formal model of *rational economic man*. According to the model (in its simplest variant), when faced with a choice, which is modeled as the selection from a set of items S, a rational decision-maker acts according to *clear preferences*, which are modeled as a partial order $\succeq$ on S. The partial order is established such that $\exists a^* \in S, \forall a \in S, a^* \succeq a$, *i.e.*, a^* is preferred over all other elements in S. a^* is the decision-maker's choice. Given another set S', such that $S \subseteq S'$, for the decision-maker's choice $a'^* \in S'$ it must hold true that $a'^* \notin S$ or $a'^* = a^*$; *i.e.*, the preference relation on S must be consistent with the preference relation on S' (see, *e.g.*, Osbourne and Rubinstein [25]). Consequently, a rational economic decision-maker can make a decision in any situation and the preferences this decision implies are consistent with the preferences implied by all previous decisions.

[5] Let us highlight that we do not introduce the so-called AGM postulates [3] here, because the *success* postulate stipulates (colloquially speaking) that "new" logical formulas are always added to the belief base and never rejected; however, we assume that, intuitively, an intelligent agent should be able to reject new beliefs under some circumstances.

[6] Less formal models of human decision-making and reasoning have been, of course, subject of in-depth study for much longer. Indeed, the management of contradictions that is at the center of this paper is also the subject of the *Shev Shema'tata*, a book on the treatment of doubt in Rabbinic law, written at the turn from the 18th to the 19th century [18].

While the model of rational economic man remains influential and is a common foundation of micro-economic curricula, its shortcomings have been criticized in high-profile scientific venues since the 1950s, notably in Herbert Simon's seminal paper *A Behavioral Model of Rational Choice* [29]. A key argument made by Simon is that the model is too simplistic in that it does not account for the information an agent has (from our perspective: the agent's *beliefs*) and hence the model can neither describe nor prescribe the real-life decision-making processes of agents or organizations. A simple example of economically irrational behavior is as follows: an agent chooses b from a set $\{b, c\}$ which implies $b \succeq c$, but chooses c from a set $\{b, c, d\}$, which implies $c \succeq b$. For instance, let us assume a choice from a set of beverages: $b := coffee, c := tea, d := juice$. After choosing coffee from the set of "tea and coffee", a rational decision-maker must not choose tea from the set of "tea, coffee, and juice", given all other things remain the same. From a knowledge representation perspective, one can of course argue that the presence of d allows us to infer something about b and/or c that makes us reverse $b \succeq c$ to $c \succeq b$[7].

Building on top of these initial insights, Tversky and Kahneman conducted a series of behavioral psychology experiments to systematically identify shortcomings of models of economic rationality that led to refined models of rational decision-making, like *prospect theory* [20], eventually winning Kahneman the Nobel Memorial Prize in Economic Sciences [19]. While a broad range of other formal models has been developed to address the aforementioned and similar shortcomings [27], further ground-breaking empirical research has emerged about other aspects of human reasoning. Most notably in the context of this paper is a line of research conducted by Jonathan Haidt (and others), showing that humans are prone to first make an intuition-based decision and, if required, then search for a "rational" (colloquially speaking) explanation [17].

To summarize, this brief overview of selected microeconomic and behavioral economics research history gives us the following insights on perspectives on human reasoning and decision-making:

1. Traditionally, humans are considered intelligent, *rational* decision-makers that act, at least roughly, according to formal model of clear and consistent preferences.
2. Empirical research about human behavior has systematically debunked assumptions about economic rationality in human decision-making, leading to a refinement of formal models of decision-making to *models of bounded rationality*.
3. More recently, additional empirical research has provided evidence for the hypothesis that humans are prone to make intuition-based decisions and then *reason backwards* to generate convincing, "rational" explanations if required.

3 Levels of Intelligent Reasoning in Face of Contradictions

From the overview of perspectives on human decision-making, we can generate three levels of intelligent reasoning in face of contradictions, which we outline in this section. In addition, we describe a fourth level that prescribes desirable behavior that – by

[7] Indeed, empirical studies (conducted decades after the publication of Simon's paper) show that humans sometimes do exactly this [7].

combining principle-based reasoning and learning perspectives – goes beyond existing perspectives on human decision-making and reasoning.

3.1 Clear Preferences

At the most primitive level, the only property one expects from a decision-maker is to be *decisive*. In microeconomic theory, this intuition is ingrained in the assumption that when observing a decision-maker who chooses one option from a set of options A, a partial order $\succeq$ on A that describes the decision-maker's preferences can be inferred, such that given the choice $a^* \in A$, it holds true that $\forall a \in A, a^* \succeq a$, *i.e.*, the decision-maker strictly prefers the choice over all other possible alternatives that could have been chosen. In its most primitive form, this model can be considered to merely cover a one-shot observation: as long as an agent is decisive, clear preferences can be inferred from a single decision and no consistency check with regard to previous decisions is performed. From a reasoning perspective, this means that an inference method must always come to a conclusion when drawing inferences from a belief base; no further conditions need to be satisfied. This one-shot approach can be compared to the behavior of a populist politician, who makes his decisions based on gut-feeling, notwithstanding that he is aware of contradicting evidence, and does not care about the long-term consistency of his actions (and speech acts).

3.2 Consistent Preferences

As an obvious next step, economists assess whether a decision-maker's preferences are consistent over a sequence of decisions; *i.e.*, given a new choice $a'^* \in A'$, such that $A \subseteq A'$, if $a'^* \in A$ then $a'^* = a^*$; this property follows from the model of clear preferences as introduced in the previous subsection (see, *e.g.*, Rubinstein [27, p. 11] for a proof). Again, from a reasoning perspective, the analogy is obvious: when drawing inferences $concl(A)$ from a belief base A, for the inferences $concl(A')$ that are drawn from a belief base A', such that $A \subseteq A'$, it must hold true that $concl(A) = concl(A')$ unless a belief in $A' \setminus A$ is accepted as an element of $concl(A')$. Consequently, we can see that the consistent preference principle is in its motivation similar to notions of "relaxed" monotony, in particular to cautious monotony [15], which can semi-formally be described as *if* $C \subseteq concl(A)$ *and* $B \subseteq concl(A)$ *then* $C \subseteq concl(A \cup B)$.

3.3 Explainable "Backwards Reasoning"

As summarized in the previous section, behavioral psychology research suggests that humans typically make intuition-based decisions and then find a "rational" explanation if necessary. This *reasoning backwards* approach has traditionally been favored by neoclassical economics, to the extent that the economist Steven Landsburg colloquializes it as follows in his best-selling popular science book *The Armchair Economist*:

"[We] stubbornly maintain the fiction that all people are rational at all times, and [...]

insist on finding rational explanations, no matter how outlandish, for all of this apparently irrational behavior[8]. " [23]

Landsburg does *not* describe observations of common human decision-making fallacies, but instead refers to – albeit with some overstatement to underline his point – a key aspect of the approach that he and some other economists use to build their models. However, in the real-world, reasoning backwards is not considered a "reasonable" approach to explain a decision or line of reasoning, which the following anecdote illustrates.

Example 1. In 2019, world-renowned association football coach José Mourinho, who at that time recently had joined Tottenham Hotspur F.C. ("the Spurs"), had the following exchange with a journalist during a press conference [9]:

- *Journalist*: "When you were at Chelsea, you were asked whether you would ever come to the Spurs and you said: 'Never, I love the Chelsea fans too much.' What has changed?"
- *Mourinho*: "[That was] before I was sacked [at Chelsea]."

From a reasoning perspective, one can say that when asked about the inconsistency between two conclusions, Mourinho produces a new belief that explains why the latter conclusion does not entail the initial conclusion. Technically, one could argue that Mourinho has successfully assured that his decision to join Tottenham is indeed consistent, because he has provided a new belief (*i.e.*, an argument) that supports his change of mind, and when considering the adoption of a belief as a part of a choice process, his preferences are consistent (*i.e.*, *economically rational*). From a logics perspective, the existence of a *conflict* between the new belief and the previous beliefs can explain why monotony of entailment is violated. However practically, it is obvious that his stated commitment to Chelsea was implied to last beyond his tenure as a coach at the club. Indeed, both Mourinho and the journalists that are present laugh about the answer; they are aware of how ridiculous the explanation that Mourinho has provided must look from the perspective of a Chelsea fan (in particular when considering that Chelsea and Tottenham are London city rivals).

3.4 Evidence-Based Principle Revision

Similarly to Tversky and Kahneman, who started off by taking formal models of economic rationality and then systematically refined them as they observed diverging human behavior in the real world, an intelligent agent should be able to start off with an *explainable* model of reasoning and decision-making and then refine it based on the observations it makes; *i.e.*, the agent should make decisions/draw inferences as follows:

1. It should employ an *explainable* formal model that prescribes and describes its behavior and satisfies some formal principles.

[8] Note that this statement precedes a defense of the approach it describes.
[9] See: `http://s.cs.umu.se/hlzdqf`

2. It should be able to refine the model and adjust its principles if it observes that changes are beneficial (based on feedback from its environment).

This hybrid approach requires a combination of symbolic and sub-symbolic (machine learning) approaches to artificial intelligence. Considering the example in the previous subsection, Mourinho could, for example, revise his reasoning principles after being subjected to the scorn of the Chelsea fans, and in the future be more conservative when discarding previously drawn conclusions, at least the ones he has publicly announced to be committed to.

4 Examples: Abstract Argumentation

Let us further illustrate the intuitions we have introduced in the previous section by providing precise formal examples. As our reasoning method, we employ abstract argumentation because it *a)* is a simple model that can be introduced without a lot of formal preliminaries and *b)* has a clear focus on managing conflicts/contradictions.

Definition 1 (Argumentation Framework [14]).
An abstract argumentation framework is a tuple $AF = (AR, AT)$, where AR is a set of arguments (propositional atoms) and $AT \subseteq AR \times AR$ is a set of attacks *between arguments in AR.*

Given an argumentation framework $AF = (AR, AT)$ and two arguments $a, b \in AR$, we say that "a attacks b" iff $(a, b) \in AT$. An argument $a \in AR$ is *acceptable* with regard to a set $S \subseteq AR$ iff for each $b \in AR$ it holds true that if b attacks a, then b is attacked by S. In abstract argumentation, key concepts are the notions of conflict-free and admissible sets.

Definition 2 (Conflict-free and Admissible Sets [14]).
Let $AF = (AR, AT)$ be an argumentation framework. A set $S \subseteq AR$ is:

- conflict-free *iff $\nexists a, b \in S$, such that a attacks b;*
- admissible *iff S is conflict-free and each argument in S is acceptable with regard to S.*

Given an argumentation framework $AF = (AR, AT)$ and a set $S \subseteq AR$, we define $S^+ = \{a \in AR | \exists b \in S,$ such that b attacks $a\}$. *Argumentation semantics* determine which sets of arguments in an argumentation framework can be considered valid conclusions. A set of such valid conclusions is called an *extension*. All argumentation semantics that have been introduced by Dung in the initial paper are based on the notion of an admissible set.

Definition 3 (Admissible Set-based Argumentation Semantics [14]).
Given an argumentation framework $AF = (AR, AT)$, an admissible set $S \subseteq AR$ is:

- *a* complete extension *iff each argument that is acceptable w.r.t. S belongs to S. $\sigma_{complete}(AF)$ returns all complete extensions of AF.*
- *a* preferred extension *of AF iff S is a maximal (w.r.t. set inclusion) admissible subset of AR. $\sigma_{preferred}(AF)$ returns all preferred extensions of AF;*

– *a* grounded extension *of AF iff S is the minimal (w.r.t. set inclusion) complete extension of AF. $\sigma_{grounded}(AF)$ returns all grounded extensions of AF.*

Given an argumentation framework $AF = (AR, AT)$ and an argumentation semantics σ, a set $S \subseteq AR$ is called a σ-extension of AF iff $S \in \sigma(AF)$. Other semantics have been defined that start of with the assumption of a maximal conflict-free (*naive*) set[10].

Definition 4 (Naive Set-based Argumentation Semantics [32]).
A conflict-free set $S \subseteq AR$ is a:

– naive extension *iff S is maximal w.r.t. set inclusion among all conflict-free sets. $\sigma_{naive}(AF)$ returns all naive extensions of AF.*
– stage extension*, iff $S \cup S^+$ is maximal w.r.t. set inclusion among all conflict-free sets,* i.e.*, there is no conflict-free set $S' \subseteq AR$, such that $(S' \cup S'^+) \supset (S \cup S^+)$. $\sigma_{stage}(AF)$ returns all stage extensions of AF.*

In the context of this paper, we are interested in how agents draw inferences from a belief base to which new beliefs are added over time. For this, we depend on the notion of argumentation framework expansion, and in particular on *normal* expansions.

Definition 5 (Argumentation Framework Expansions [8]).
An argumentation framework $AF' = (AR', AT')$ is:

– *an **expansion** of another argumentation framework $AF = (AR, AT)$ (denoted by $AF \preceq_E AF'$) iff $AR \subseteq AR'$ and $AT \subseteq AT'$.*
– *a **normal expansion** of an argumentation framework $AF = (AR, AT)$ (denoted by $AF \preceq_N AF'$) iff $AF \preceq_E AF'$ and $\nexists (a, b) \in AT' \setminus AT$, such that $a \in AR \wedge b \in AR$.*

Colloquially speaking, a normal expansion of an argumentation framework adds new arguments to the argumentation framework, but neither removes arguments nor changes attacks between existing arguments. To support the design and analysis of argumentation semantics, formal argumentation principles have been defined [30,5]. For example, the *uniqueness* principle stipulates that an argumentation semantics must return exactly one extension, given any argumentation framework.

4.1 Clear Preferences

From an argumentation perspective, an agent has clear preferences iff it can reach an unambiguous conclusion, given any argumentation framework and the argumentation semantics it employs. We can illustrate perspectives on this property given a *particular* argumentation framework, *e.g.*, $AF = (AR, AT) = (\{a, b, c\}, \{(a, a), (b, c), (c, b)\})$. Figure 1 depicts the argumentation framework. Below are some examples of how different argumentation semantics resolve AF:

– Stable semantics: $\sigma_{stable}(AF) = \{\}$;

[10] More semantics exist, some of which address well-known issues with the semantics whose definitions we provide in this paper. However, we consider an in-depth overview of argumentation semantics out-of-scope.

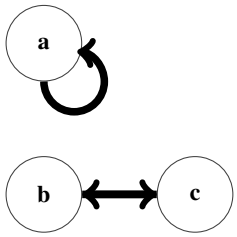


Fig. 1: Given stable semantics, self-contradicting arguments may lead to the inability to reach any conclusion, *e.g.*, $\sigma_{stable}(\{a, b, c\}, \{(a, a), (b, c), (c, b)\}) = \{\}$.

- Grounded semantics: $\sigma_{grounded}(AF) = \{\{\}\}$;
- Preferred semantics: $\sigma_{preferred}(AF) = \{\{b\}, \{c\}\}$.

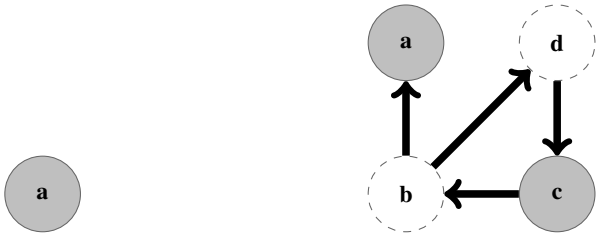
It is obvious that stable semantics does not satisfy the notion of *clear preferences*: it does not return any extension for our argumentation framework. Conversely, preferred semantics returns the extensions $\{a\}$ and $\{b\}$. This does not reflect the clear preferences principle, either, because several extensions are returned. However, an intelligent agent that employs the semantics can certainly come to a decisive conclusion, for example by considering use case-specific meta-data (like a time-stamp or the source of an argument), or by simply *breaking the tie* with an arbitrary method that considers language-specific properties, like identifiers of the arguments[11]. Consequently, we argue that it depends on the exact application scenario whether one wants an argumentation semantics to be uniquely defined or not. For example, in one legal reasoning scenario, it can make sense to dismiss conflicting statements of two witnesses as mutually inconsistent, while in another scenario, it can be better to consider both statements and then select a preferred statement based on situational context or meta-data (which is aligned with the concept of *burden of persuasion*, see Prakken and Sartor [26]).

4.2 Consistent Preferences

To align with the *consistent preferences* property of economic rationality, we can create a straight-forward argumentation principle (see our ongoing line of work [22,21][12]): we assume that an agent, given an argumentation semantics σ, resolves an argumentation framework $AF = (AR, AT)$ by selecting any σ-extension E of AF ($E \in \sigma(AF)$). This selection implies the preferences $\forall S \in 2^{AR}, E \succeq S$. When continuing the interaction with its environment, the agent adopts new, and potentially conflicting beliefs, *i.e.*, it normally expands AF and creates $AF' = (AR', AT'), AF \preceq_N AF'$. When determining the σ-extensions of AF', the agent must find at least on extension ($\exists E' \in \sigma(AF)$), such that the preferences implied by E' and AF' ($\forall S' \subseteq 2^{AR'}, E' \succeq S$) are consistent with the preferences implied by E and AF. Figure 2 illustrates the concept of consistent preferences in abstract argumentation. For example, let us assume argument a denotes that a new business strategy should be executed, to which

[11] Note that this would be a violation of the *language independence* principle.

[12] In these works, we name the principle *weak reference independence*.



(a) $AF = (AR, AT) = (\{a\}, \{\})$.

(b) $AF' = (AR', AT') = (\{a, b, c, d\}, \{(b, a), (c, b), (d, c), (b, d)\})$.

Fig. 2: Consistent preferences. Assuming stage semantics, we have $\sigma_{stage}(AF) = \{\{a\}\}$ and $\sigma_{stage}(AF') = \{\{a, c\}, \{a, d\}, \{b\}\}$. All σ_{stage}-extensions of AF' imply consistent preferences with regard to the only σ_{stage}-extension of AF. In contrast, assuming preferred semantics, we have $\sigma_{preferred}(AF) = \{\{a\}\}$ and $\sigma_{preferred}(AF') = \{\{\}\}$; the only $\sigma_{preferred}$-extension of AF implies the preferences $\forall S \in 2^{AR}, \{a\} \succeq S$, which is inconsistent with the preferences implied by the only $\sigma_{preferred}$-extension of AF: $\forall S' \in 2^{AR'}, \{\} \succeq S'$.

an agent first commits: $AF = (AR, AT) = (\{a\}, \{\})$, from which we obviously conclude $\{a\}$. However, by consulting multiple stakeholders, the agent collects the additional arguments b, c, and d that directly or indirectly argue for or against the strategy: $AF' = (AR', AT') = (\{a, b, c, d\}, \{(b, a), (c, b), (d, c), (b, d)\})$. Now, considering some argumentation semantics, for example preferred semantics, the only conclusion we can draw from AF' is $\{\}$ (the only extension/valid conclusion does not contain any arguments); this implies the preference $\{\} \succeq \{a\}$, which is inconsistent with the preference $\{a\} \succeq \{\}$ as implied by the previous decision. In contrast, some other semantics, such as stage semantics, do not imply inconsistent preferences *in this scenario*[13]. $\sigma_{stage}(AF) = \{\{a\}\}$ and $\sigma_{stage}(AF') = \sigma_{stage}(AF') = \{\{a, c\}, \{a, d\}, \{b\}\}$: because all σ_{stage}-extensions of AF' include an argument that is not in AR, the preferences implied by selecting any of the extensions are obviously consistent with the preferences implied by inferring $\{a\}$ from AF.

Let us note that an open question, which we touch upon in Section 5, is how to adjust the consistent preferences principle to account for "undecided" arguments, *i.e.*, arguments that are, given an extension, neither part of the extension nor attacked by any argument in the extension. Also, similar argumentation principles that are based on other well-known properties can be and have been introduced, for example an abstract argumentation equivalent of restricted (cautious) monotony [21,22].

4.3 Explainable "Backwards Reasoning"

An important feature of an intelligent agent is the ability to explain its inferences and the resulting actions. Indeed, economists who build formal models of human decision-

[13] Let us note that stage semantics does not generally imply consistent preferences, given any argumentation framework and any of its normal expansions, see [22].

making typically do not claim that their models are accurate representations of what goes on in a human's mind, but instead argue that when observing a human decision-maker, their models are sufficiently precise to describe the decision-maker's behavior in a explainable (that is: formally analyzable) manner. In the artificial intelligence community, the design and analysis of *explainable* agents is a research direction that has gained tremendous traction over the past years [4]. Agents that employ symbolic approaches to automated reasoning, such as abstract argumentation, are typically considered *explainable*, because each inference and action can be linked to the formal model that generated it (see, *e.g.*, Zhong *et al.* [33]). However, when considering the iterative argumentation approach we take in the context of this paper, it is clear that merely pointing out general semantics behavior is not always sufficient to explain why exactly the inferences drawn from an argumentation framework are fundamentally different than the inferences that are subsequently drawn from one of its (normal) expansions. To some extent, merely explaining an inference process by pointing to the entire formal model that has been used to infer it resembles the *reasoning backwards* approach as introduced as a description of human reasoning in behavioral economics. From an argumentation perspective, we argue that an agent can take two approaches to reasoning backwards:

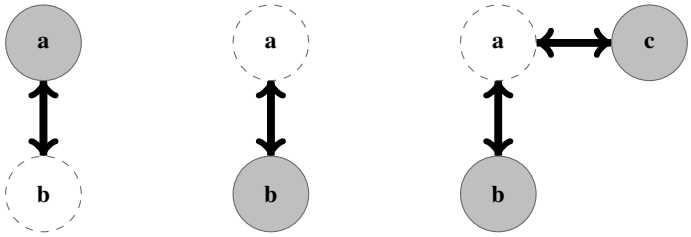
1. It can take a principle that happens to be satisfied to explain the result of its inference process.
2. If asked why a specific principle is violated, it can generate arguments and add them to the argumentation framework, so that the principle is no longer violated.

As mentioned above, the first approach is obvious, and reflected in the way *explainable* argumentation is typically presented. The second approach reflects the *Mourinho* example (Example 1), which is illustrated as a sequence of argumentation frameworks by Figure 3:

1. We start with an initial argumentation framework $AF = (\{a, b\}, \{(a, b), (b, a)\})$. a denotes the obligation of maintaining the respect of the Chelsea fans while b denotes taking a job at Tottenham; a and b attack each other. Our agent (Mourinho) infers a, deciding to stay committed to Chelsea.
2. Later, our agent has a change of mind, and instead infers $\{b\}$ from $AF' = AF$ and takes a job at Tottenham.
3. Another agent (the journalist) scrutinizes the Mourinho agent by highlighting that the inference process implies inconsistent preferences.
4. The Mourinho agent responds to the scrutiny by producing an argument c (the relief of the loyalty obligation because Chelsea has sacked him), which is in mutual conflict with argument a. Note that inferring $\{b, c\}$ from AF'' does not imply preferences that are inconsistent with the preferences implied by inferring $\{a\}$ from AF.

4.4 Evidence-Based Principle Revision

Let us go back to the previous example (Figure 3). However, we now assume the *Mourinho* agent is using the relational principle we have semi-formally introduced



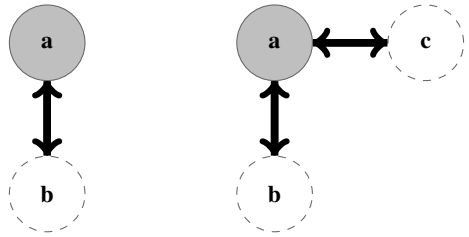
(a) AF. (b) AF'. (c) AF''.

Fig. 3: Reasoning backwards. Given the argumentation framework $(\{a, b\}, \{(a, b), (b, a)\})$, an agent first concludes $\{a\}$and at a later stage concludes $\{b\}$. When asked about the reason for the inconsistency (preference reversal), the agent produces argument c (generating AF'') that restores consistent preferences.

in Subsection 4.2 to ensure consistent preferences. In the example, this means that the agent must not infer $\{b\}$ from AF' after having inferred $\{a\}$ from AF; the expansion to AF'' is required to then infer $\{b, c\}$, which is a principle-compliant conclusion. Let us assume that after drawing this inference (and joining Tottenham), our agent's reputation is severely damaged, which makes the agent reflect about its inference process. Satisfying the consistent preferences principle may have been a reasonable starting point, but we want to be able to further evolve from there. Ideally, the agent analyzes its own inference process and searches for principle-based improvements it can make. In our example, the agent can, for instance assuming that it is using stage semantics, observe that the semantics also supports inferring $\{a\}$ from AF'': *i.e.*, $\sigma_{stage}(AF'') = \{\{a\}, \{b, c\}\}$. Consequently, the agent can "learn" a new principle that stipulates the following: given two argumentation frameworks AF^* and AF^{**} and a conclusion E^* of AF^* ($E^* \in \sigma_{stage}(AF^*)$), iff inferring a conclusion E' from AF^{**} ($E^{**} \in \sigma_{stage}(AF^{**})$) is possible such that $E^* \subseteq E^{**}$, do not infer a conclusion D^{**} from AF^{**} such that $E^* \not\subseteq D^{**}$. The agent can apply this principle and draw inferences in future scenarios accordingly (depicted by Figure 4). However, first the agent would need to (formally) verify whether enforcing this new principle implies a violation of any other principle that the agent has already adopted (in our example, the agent may still want to satisfy the consistent preferences principle), and if so, whether previously adopted principles should be relaxed or entirely discarded.

5 Research Directions

Based on the position we establish in the previous sections, we provide an overview of relevant ongoing research directions and highlight open challenges. Again, our focus is on formal argumentation as an exemplary method for automated non-monotonic reasoning.



(a) AF^*. (b) AF^{**}.

Fig. 4: Evidence-based principle revision. Let us assume our agent has received negative feedback from its actions that were based on the inferences drawn in Figure 3. To learn from this experience, the agent adjusts its reasoning principles and now always keeps previously inferred conclusions (arguments) to the extent its semantics supports this.

5.1 Consistent Preference and Undecided Beliefs

Some of the argumentation examples we present in this paper draw from ongoing research on economic rationality and formal argumentation [22,21]. An open question in this line of research is how to adjust the model of consistent preference in abstract argumentation to support the notion of *undecided* arguments[14]. Let us highlight that this question cannot be addressed by straight-forward tweaks of the economic rationality-based argumentation principle, in particular because an agent must eventually commit to a course of action; *i.e.*, some arguments must not remain undecided. This can be illustrated with the help of a simple example. We have two weather report sources: one reports that it will rain (argument r) and the other reports it will not rain (argument $\neg r$). Obviously, r and $\neg r$ attack each other. We want to decide whether to take an umbrella with us (argument u). If we think it does not rain, we do not take an umbrella with us ($\neg r$ attacks u). Figure 5 depicts the corresponding argumentation framework.

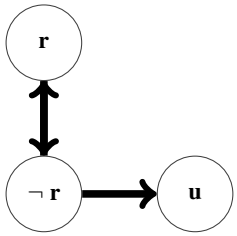


Fig. 5: $AF = (\{r, \neg r, u\}, \{(r, \neg r), (\neg r, r), (\neg r, u)\})$. How can we manage undecided arguments if we cannot be undecided about actions?

[14] Given an argumentation framework and a semantics' extension of this framework, the undecided arguments are all arguments that are neither in the extension, nor attacked by any of the arguments in the extension.

Given, for example, grounded semantics $\sigma_{grounded}$, all arguments in the argumentation framework are undecided. However, we must eventually make a decision on whether or not to take the umbrella with us; *i.e.*, to support undecided arguments, we need to define two argument types: *belief arguments* that may be undecided and *action arguments* that must never be undecided.

5.2 Burdens of Persuasion

When analyzing consistency and monotony properties of inference methods like formal argumentation approaches, it can be useful to apply intuitions that are provided by well-established practical research domains. In this regard, a particularly interesting concept is the notion of the *burden of persuasion* in legal research and practice. In case of two conflicting statements, the burden of persuasion can be placed on one of the statements, which implies that this statement requires additional justification; otherwise, it will be automatically defeated. For example, given two contradicting witness statements, of which one provides an alibi of the defendant, whereas the other one claims the defendant was at the crime scene at the time of the crime, the burden of persuasion could be laid on the latter argument to reflect the notion of *in dubio pro reo*[15]. Models of burdens of persuasion have already been introduced to formal argumentation approaches [26,10]. In these approaches, the burden of persuasion is explicitly modeled. In contrast, from the perspective of *consistent* inference, the burden of persuasion can automatically be placed on new arguments when expanding an argumentation framework; *i.e.*, if considering a new argument as part of the conclusion violates a consistency/monotony property (because the new argument is, directly or indirectly, in conflict with an argument that is part of a previous conclusion), the burden of persuasion is placed on this argument; additional conditions must be satisfied to allow for this argument to "kick out" the previously inferred argument[16]. Formally integrating this intuition with models of burdens of persuasion and consistency/monotony properties of formal argumentation can be considered promising future research.

5.3 Intuitive Rationality

Independently of the research on formal models of economic rationality and formal argumentation, recent research has started to shed light on what humans intuitively think are "reasonable" conclusions that can be drawn from argumentation frameworks [11,12]. The results suggest that while there is not necessarily one semantics whose behavior is more intuitive to most humans than all other semantics, some semantics (notably grounded and CF2 semantics[17]) seem to exhibit particularly intuitive behavior. As a result of these studies, SCF2 semantics has been introduced, which addresses some issues CF2 semantics has with regard to the handling of self-attacking arguments and

[15] This is a constructed example that does not fully reflect real-world legal reasoning.

[16] This notion is reflected by *loop-busting* approaches that have been proposed in the context of formal argumentation and that are based on Talmudic logic [2].

[17] For the sake of conciseness we do not introduce CF2 semantics in this paper; the semantics is introduced by Baroni *et al.* in [6].

even cycles that exceed a certain length [13]. The studies shed some light on human evaluations of argumentation principles, which can, however, be investigated more comprehensively. In particular, it is worth examining how well intuitive human assessments align with the consistent preference argumentation principle that is based on economic rationality (see Subsection 4.2), as well as with other principles that can emerge from cross-disciplinary perspectives on "rational" and "consistent" reasoning and decision-making.

5.4 Neuro-Symbolic Artificial Intelligence

Recently, combining machine learning and symbolic reasoning approaches has re-emerged as a hot topic in artificial intelligence research [16]. This trend is possibly accelerated because the machine learning break-throughs of the last decade have created the initial expectation of rapid and continuous progress, which machine learning alone cannot live up to. However, the integration of machine learning approaches and symbolic methods (which is sometimes referred to as *neural-symbolic AI*) has been a well-established research direction since several decades [1]. In Subsection 4.4, we illustrate by example that a neuro-symbolic AI approach can be considered promising to allow for the evidence-based revision of reasoning (argumentation) principles. While formal argumentation has been integrated with machine learning methods, in particular in the context of argument mining [9], to our knowledge no research combines these hybrid approaches with a principle-based perspective.

To realize our proposal of an agent that can learn reasoning principles as it observes and interacts with its environment, we need create formal models and implementations at the intersection of non-monotonic symbolic reasoning and reinforcement learning, to find answers to the following questions. *i)* Which principles should an agent inhibit statically by design and which principles should be learnable? *ii)* How can we design principles that allow for a parameterization that facilitates learning? *iii)* To what extent is principle revision use-case agnostic, to what extent is it use-case-dependent? *iv)* When an agent learns new principles and hence updates its inference method, how does it trade-off consistency with regard to previously drawn inferences and compliance with the newly learned principles?

6 Conclusion

In this paper, we have introduced a formal perspective that takes inspirations from models of human models of decision-making reasoning to define levels of intelligent reasoning, *i.e.*, the ability of an agent to:

1. reason in face of contradictions;
2. reason according to well-established principles, like the *clear and consistent preferences* principle that follows from economic rationality;
3. explain the resolution of contradictions according to whatever reasoning principles that are satisfied in a given scenario;
4. dynamically revise a principle-based inference process based on feedback the agent perceives as the result of interactions with its environment.

This perspective integrates well with a long-running line of research on non-monotonic reasoning approaches, which we have illustrated for formal (abstract) argumentation. In particular, *dynamic* models of formal argumentation that cover the expansion and iterative resolution of argumentation frameworks, considering fundamental properties of non-monotonic reasoning. However, as outlined in this paper, these models need further refinement to fully reflect the idea of explainable intelligent reasoning in face of contradictions.

Acknowledgments. The authors thank Amro Najjar, Michele Persiani, and the anonymous reviewers for their useful feedback. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

1. Neural-Symbolic Learning Systems, pp. 35–54. Springer Berlin Heidelberg, Berlin, Heidelberg (2009), https://doi.org/10.1007/978-3-540-73246-4_4
2. Abraham, M., Gabbay, D.M., Schild, U.J.: The handling of loops in talmudic logic, with application to odd and even loops in argumentation. HOWARD-60: A Festschrift on the Occasion of Howard Barringer's 60th Birthday (2014)
3. Alchourrón, C.E., Gärdenfors, P., Makinson, D.: On the logic of theory change: Partial meet contraction and revision functions. The Journal of Symbolic Logic **50**(2), 510–530 (1985)
4. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. p. 1078–1088. AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2019)
5. Baroni, P., Giacomin, M.: On principle-based evaluation of extension-based argumentation semantics. Artificial Intelligence **171**(10), 675 – 700 (2007). https://doi.org/https://doi.org/10.1016/j.artint.2007.04.004, http://www.sciencedirect.com/science/article/pii/S0004370207000744, argumentation in Artificial Intelligence
6. Baroni, P., Giacomin, M., Guida, G.: Scc-recursiveness: a general schema for argumentation semantics. Artificial Intelligence **168**(1), 162 – 210 (2005). https://doi.org/https://doi.org/10.1016/j.artint.2005.05.006
7. Bateman, I., Munro, A., Rhodes, B., Starmer, C., Sugden, R.: A test of the theory of reference-dependent preferences. The quarterly journal of economics **112**(2), 479–505 (1997)
8. Baumann, R., Brewka, G.: Expanding argumentation frameworks: Enforcing and monotonicity results. COMMA **10**, 75–86 (2010)
9. Cabrio, E., Villata, S.: Five years of argument mining: A data-driven analysis. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. p. 5427–5433. IJCAI'18, AAAI Press (2018)
10. Calegari, R., Sartor, G.: Burden of persuasion in argumentation. arXiv preprint arXiv:2009.10244 (2020)
11. Cramer, M., Guillaume, M.: Empirical cognitive study on abstract argumentation semantics. Frontiers in Artificial Intelligence and Applications (2018)
12. Cramer, M., Guillaume, M.: Empirical study on human evaluation of complex argumentation frameworks. In: Calimeri, F., Leone, N., Manna, M. (eds.) Logics in Artificial Intelligence. pp. 102–115. Springer International Publishing, Cham (2019)

13. Cramer, M., van der Torre, L.: Scf2-an argumentation semantics for rational human judgments on argument acceptability. In: Proceedings of the 8th Workshop on Dynamics of Knowledge and Belief (DKB-2019) and the 7th Workshop KI\& Kognition (KIK-2019) co-located with 44nd German Conference on Artificial Intelligence (KI 2019), Kassel, Germany, September 23, 2019. pp. 24–35 (2019)
14. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence **77**(2), 321–357 (1995)
15. Gabbay, D.M.: Theoretical foundations for non-monotonic reasoning in expert systems. In: Apt, K.R. (ed.) Logics and Models of Concurrent Systems. pp. 439–457. Springer Berlin Heidelberg, Berlin, Heidelberg (1985)
16. Geffner, H.: Model-free, model-based, and general intelligence. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. p. 10–17. IJCAI'18, AAAI Press (2018)
17. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychological review **108**(4), 814 (2001)
18. Jacobs, L.: Rabbi aryeh laib heller's theological introduction to his "shev shema'tata". Modern Judaism **1**(2), 184–216 (1981), http://www.jstor.org/stable/1396060
19. Kahneman, D.: Maps of bounded rationality: Psychology for behavioral economics. American economic review **93**(5), 1449–1475 (2003)
20. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. Econometrica **47**(2), 263–291 (1979)
21. Kampik, T., Gabbay, D.: Towards diarg: An argumentation-based dialogue reasoning engine. In: SAFA@ COMMA. pp. 14–21 (2020)
22. Kampik, T., Nieves, J.C.: Abstract argumentation and the rational man. Journal of Logic and Computation **31**(2), 654–699 (02 2021). https://doi.org/10.1093/logcom/exab003, https://doi.org/10.1093/logcom/exab003
23. Landsburg, S.: The Armchair Economist (revised and updated May 2012): Economics & Everyday Life. Free Press (2007)
24. Lehmann, D., Magidor, M.: What does a conditional knowledge base entail? Artificial Intelligence **55**(1), 1 – 60 (1992), http://www.sciencedirect.com/science/article/pii/000437029290041U
25. Osborne, M.J., Rubinstein, A.: Models in Microeconomic Theory. Open Book Publishers (2020). https://doi.org/10.11647/OBP.0204
26. Prakken, H., Sartor, G.: A logical analysis of burdens of proof. Legal evidence and proof: Statistics, stories, logic pp. 223–253 (2009)
27. Rubinstein, A.: Modeling bounded rationality. MIT press (1998)
28. Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.C., Flammini, A., Menczer, F.: The spread of low-credibility content by social bots. Nature communications **9**(1), 1–9 (2018)
29. Simon, H.A.: A Behavioral Model of Rational Choice. The Quarterly Journal of Economics **69**(1), 99–118 (02 1955). https://doi.org/10.2307/1884852
30. van der Torre, L., Vesic, S.: The principle-based approach to abstract argumentation semantics. IfCoLog Journal of Logics and Their Applications **4**(8) (October 2017)
31. Turing, A.M.: Computing Machinery and Intelligence, pp. 23–65. Springer Netherlands, Dordrecht (2009), https://doi.org/10.1007/978-1-4020-6710-5_3
32. Verheij, B.: Two approaches to dialectical argumentation: admissible sets and argumentation stages. Proc. NAIC **96**, 357–368 (1996)
33. Zhong, Q., Fan, X., Luo, X., Toni, F.: An explainable multi-attribute decision model based on argumentation. Expert Systems with Applications **117**, 42 – 61 (2019). https://doi.org/https://doi.org/10.1016/j.eswa.2018.09.038, http://www.sciencedirect.com/science/article/pii/S0957417418306158