From User Preferences to Base Score Extraction Functions in Gradual Argumentation

Aniol Civit
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Barcelona, Spain
acivit@iri.upc.edu

Antonio Rago
King's College London
London, United Kingdom
antonio.rago@kcl.ac.uk

Antonio Andriella
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Barcelona, Spain
aandriella@iri.upc.edu

Guillem Alenyà
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Barcelona, Spain
galenya@iri.upc.edu

Francesca Toni
Imperial College London
London, United Kingdom
ft@imperial.ac.uk

ABSTRACT

Gradual argumentation is a sub-field of Computational Argumentation from symbolic AI which is attracting attention for its ability to support transparent and contestable AI systems. It is considered a useful tool in domains such as decision-making, recommendation, debate analysis, amongst others. The outcomes in such domains are usually dependent on the arguments' base scores, which must be selected carefully. Often, this selection process requires user expertise and may not always be straightforward. On the other hand, organising the arguments by preference could simplify the task. In this work, we introduce *Base Score Extraction Functions*, which provide a mapping from users' preferences over arguments to base scores. These functions can be applied to the arguments of a *Bipolar Argumentation Framework* (BAF), supplemented with preferences, to obtain a *Quantitative Bipolar Argumentation Framework* (QBAF), allowing the use of well-established computational tools in gradual argumentation. We outline the desirable properties of Base Score Extraction Functions, discuss some design choices, and provide an algorithm for base score extraction. Our method incorporates an approximation of non-linearities in human preferences to allow for better approximation of the real ones. Finally, we evaluate our approach both theoretically and experimentally in a robotics setting, and offer recommendations for selecting appropriate gradual semantics in practice.

KEYWORDS

Gradual Argumentation; Base Score Extraction; User Preferences

ACM Reference Format:
Aniol Civit, Antonio Rago, Antonio Andriella, Guillem Alenyà, and Francesca Toni. 2026. From User Preferences to Base Score Extraction Functions in Gradual Argumentation. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/LIEI2830

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/LIEI2830

1 INTRODUCTION

Abstract Argumentation Frameworks (AFs) [20] have long been known to be a promising method for adaptable decision-making systems [2]. These frameworks are composed of arguments and relations between them, representing conflicts. Semantics, whether extension-based [20] or gradual [19], may then be used to determine the acceptance of the arguments, e.g., as potential decisions to be taken. However, in real-world scenarios involving human reasoning, having only arguments and their relations is rarely sufficient to align the decision towards human intentions. For example, some studies claim that the acceptability of arguments is related to the preferences of the audience to which they are addressed [23].

Aligning decision-making systems with human preferences [22], and also values [39], has never been more important, given their increasing prevalence in humans' daily lives. Within the field of Computational Argumentation, this has been accomplished by extending AFs into Preference-based AFs (PAFs) [5], in which the arguments' acceptability is based not only on relations and the chosen semantics, but also on a preference ordering.

A current limitation in extension-based PAFs is that preferences are used to reduce or modify the framework [7, 23]. Those reductions correspond to different intuitions, and each one is subject to criticism. Moreover, many decision-making settings require a single output, whereas qualitative frameworks may return multiple undominated options or none at all [6]. In contrast, Gradual Argumentation ensures a determinate outcome while remaining transparent about score contributions [15]. In [12], reductions are avoided by extending PAFs to *Quantitative Bipolar AFs* (QBAFs) [10]. In QBAFs, the arguments can attack and support each other, and each argument is assigned a base score, i.e. an intrinsic strength. Gradual semantics of QBAFs [10] then assign a final strength to each argument, providing a different form of information than traditional AFs. QBAFs are currently used in several domains, such as decision support [12], product recommendation [32], review aggregation [17, 31], and decision-making [21], given their potential for giving transparent and contestable systems [25, 43]. Nonetheless, in some scenarios, it is often unclear how to set the arguments' base scores.

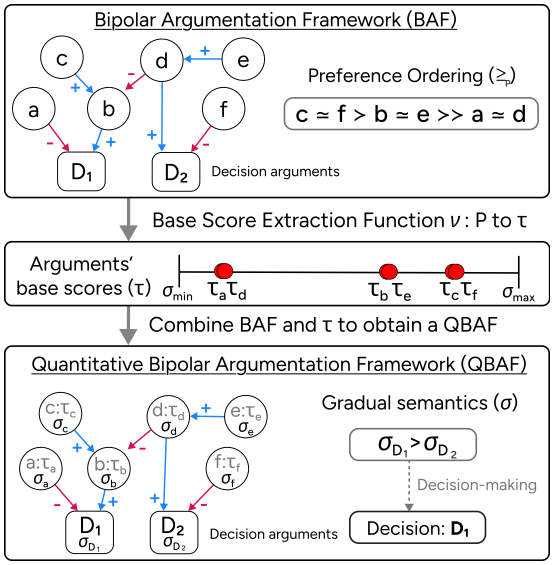

Figure 1: We introduce a methodology for converting a BAF supplemented with a preference ordering of the arguments (top), into a QBAF (bottom). A Base Score Extraction Function ν is applied to the preference ordering to obtain the arguments' base scores τ (middle), which allow for the QBAF to be used as normal for decision-making. The output of the decision-making system is adapted to the user's preferences.

In this work, we propose a theoretically and experimentally motivated methodology for generating these base scores based on a given set of user preferences. Our approach is based on adding preferences to *Bipolar Argumentation Frameworks* (BAFs) [14], i.e., AFs with an additional relation of support (but without a base score). We operate on the hypothesis that the given preference ordering allows for a Base Score Extraction Function (BSEF) to predict the arguments' base scores. Our approach is illustrated in Fig. 1. Note that, to ensure that it is able to represent realistic user preferences, on top of the usual 'I prefer one argument over another', we introduce the possibility of having a preference relation in which one argument is *much more* preferred than another. We then give some general axioms that any BSEF should satisfy. Next, we provide different design choices for adapting the base score extraction, which are formally characterised as potentially desirable, optional properties. Finally, we introduce some concrete BSEFs and evaluate them through both a theoretical analysis and experimentation on a proposed use case, namely *assistive feeding in robotics*.

Our contributions are as follows: (i) the introduction of BSEFs, which extract the base scores of a set of arguments given a preference ordering, extended to non-linear preferences; (ii) the definition of a set of axioms and properties for BSEFs; (iii) the introduction of two concrete BSEFs; and (iv) a theoretical and experimental evaluation of the two BSEFs.

To foster transparency and reproducibility, we have released the full implementation at https://github.com/acivit/From-Preferences-To-Base-Score-Extraction-Functions.

2 PRELIMINARIES

In this section, we recall the definitions of BAFs, QBAFs and their gradual semantics.

A BAF [14] is a triple $\mathcal{B} = \langle \mathcal{X}, \mathcal{R}^-, \mathcal{R}^+ \rangle$, consisting of a finite set of arguments $\mathcal{X}$, a binary relation of attack $\mathcal{R}^- \subseteq \mathcal{X} \times \mathcal{X}$, and a binary relation of support $\mathcal{R}^+ \subseteq \mathcal{X} \times \mathcal{X}$.

A QBAF [10] is a quadruple $\mathcal{Q} = \langle \mathcal{X}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$, where $\langle \mathcal{X}, \mathcal{R}^-, \mathcal{R}^+ \rangle$ is a BAF and $\tau : \mathcal{X} \to [0, 1]$ is a base score function. For any $a \in \mathcal{X}$, $\tau(a)$ is the base score of a.

In this work, we use the standard range for both the base scores and argument strengths, i.e., $[0, 1]$, though others exist [4]. The strength of an argument $a \in \mathcal{X}$ is given by $\sigma(a)$, where $\sigma : \mathcal{X} \to [0, 1]$ is a gradual semantics [10].

As mentioned, we use QBAFs for decision-making. In this context, the possible options of the decision process are included in the QBAF. These are called decision arguments. The QBAF is represented as a set of trees, with the root of each tree corresponding to a decision argument [33]. (We leave cyclic QBAFs to future work, as in [30, 31, 33].) Let $\mathcal{Q}$ be a QBAF $\langle \mathcal{X}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$. For any arguments $a, b \in \mathcal{X}$, let a path from a to b be defined as a sequence $(\gamma_0, \gamma_1), ..., (\gamma_{n-1}, \gamma_n)$ of length $n > 0$, where $\gamma_0 = a$ and $\gamma_n = b$, and, for any $1 \leq i \leq n, (\gamma_{i-1}, \gamma_i) \in \mathcal{R}^+ \cup \mathcal{R}^-$. Then, given a set of decision arguments $D \subseteq \mathcal{X}$, $\mathcal{Q}$ is a QBAF for D iff i) $\nexists a \in \mathcal{X} \setminus \{D\}$ such that $\exists d \in D$ where $(d, a) \in \mathcal{R}^+ \cup \mathcal{R}^-$ ii) $\forall a \in \mathcal{X} \setminus \{D\}$ there is a path from a to at least one $d \in D$; and iii) $\nexists a \in \mathcal{X}$ with a path from a to a. We set the base score of the decision arguments to 0.5.

In the remainder of the paper, unless otherwise specified, we will assume a BAF $\langle \mathcal{X}, \mathcal{R}^-, \mathcal{R}^+ \rangle$.

As a running example, we will use robot-assisted feeding. Here, a robot feeds a user who cannot eat independently and must decide the pace of feeding, either slow or fast. The robot uses a BAF drawing arguments from both the user and the robot's observations, e.g., user status and history. The arguments are: (a) Eating slowly causes boredom; (b) eating slowly does not stress the patient out; (c) the patient is vulnerable and stress must be avoided; (d) eating slowly reduces time with a visiting niece; (e) the patient wants to tell their niece something important, and doing so will help the patient to feel more relaxed; and (f) eating quickly carries a (low) risk of dysphagia. The decision arguments are *slow* (D_1) and *fast* (D_2). The BAF is represented in Fig. 1. The aim is to select the best option, namely any decision argument with the highest final strength in the QBAF obtained after predicting the arguments' base scores to adhere to the preferences (again, as in Fig. 1).

3 EXTRACTING BASE SCORES FROM PREFERENCE ORDERINGS

In this section, we show that adding preferences to BAFs allows us to define a BSEF, which may exhibit some defined desirable axioms and properties. Then, the BSEF is used to extract the arguments' base scores, effectively converting the BAF into a QBAF.

3.1 Adding Preferences to BAFs

Adding preferences over arguments is a common approach to personalising argumentation frameworks for users [5, 26, 27]. We formally define preferences over arguments as follows:

Definition 1. *A preference $\succeq$ over $\mathcal{X}$ is a reflexive and transitive relation on $\mathcal{X}$. Given $a, b \in \mathcal{X}$, $a \succeq b$ and $a \not\preceq b$ will be denoted as $a \succ b$, while $a \succeq b$ and $a \preceq b$ will be denoted as $a \simeq b$.*

Setting the base scores according to the preference ordering of a BAF's arguments in a decision-making environment can give an intuition of the most suitable decisions in line with a user's preferences. The following examples show the importance of properly setting the arguments' base scores:

Example 1. Consider the framework from Fig. 1. If the base scores are arbitrarily set to 0.5, giving equal importance to all arguments, using the Quadratic Energy (QE) Model [29], the final strengths of the decision arguments are $\sigma(D_1) = 0.5$ for slow, and $\sigma(D_2) = 0.505$ for fast. Then the robot would choose to move fast.

Example 2. (Example 1 Cont.) Imagine a scenario where a user gives more importance to the arguments related to its safety (e.g., c and f), less importance to those related to its relaxation (e.g., b and e), and the least importance to the arguments related to its enjoyment (e.g., a and d). The resulting preference ordering is: $c \simeq f \succ b \simeq e \succ a \simeq d$. Intuitively, the base scores of c and f should be greater than those of b and e, and which should in turn be greater than the base scores of a and d. For example, they could be set to $\tau(c) = \tau(f) = 0.75$, $\tau(b) = \tau(e) = 0.5$, and $\tau(a) = \tau(d) = 0.25$. Here, using the QE Model again, the final strengths are $\sigma(D_1) = 0.54$ and $\sigma(D_2) = 0.45$. Then the robot would choose to move slowly, which is the opposite of what was selected earlier.

3.2 Adapting to Non-Linear Human Preferences

Human preference relations are generally not linear [28]. People's choices and priorities often fluctuate instead of increasing or decreasing in a simple, linear manner as conditions change. This non-linearity makes it complex to set the base scores properly. A possible approach to including these non-linearities could be achieved by allowing the user to order their preferences gradually within a range. However, setting a gradual score is usually slower and has more sample noise than cardinal orderings [38, 42].

In this work, we approximate the non-linearities by introducing a new preference relation, where a user may have a *much greater* ($\succ\!\!\succ$) preference for one argument over another. This relation considers the non-linearity and is easy to recognise from human feedback. Definition 1 is extended as follows:

Definition 2. *Given $a, b \in \mathcal{X}$, if a is much more preferred than b, the preference relation will be denoted as $a \succ\!\!\succ b$. This relation is transitive, and $a \succ\!\!\succ b$ implies $a \succ b$. Therefore, a preference relation between arguments $a, b \in \mathcal{X}$ in a preference ordering $\succeq$ can be:*

- *$a \simeq b$, which denotes indifference;*
- *$a \succ b$, which denotes strict preference;*
- *$a \succ\!\!\succ b$, which denotes a much stronger preference.*

Example 3. Given the arguments $a, b, c \in \mathcal{X}$, with a preference ordering: $a \succ\!\!\succ b \succ c$, we may expect that $\tau(a)$ is *much greater* than $\tau(b)$ and $\tau(c)$, e.g., 0.9, 0.3, and 0.1, respectively.

This new relation breaks the linearity in the preference ordering, providing a more realistic approximation to human preferences, since it has a greater expressiveness.

3.3 Base Score Extraction Functions

Although PAFs and BAFs offer means to capture the evaluation of arguments, they remain categorical and context-sensitive. In contrast, integrating a method to extract base scores from preference orderings would provide a more granular and objective evaluation mechanism. Furthermore, the literature on QBAFs offers methods to provide transparent and counterfactual explanations [24, 44] and allows for fine-tuning the framework to better adapt to users, which is crucial in some decision-making systems, such as robotics.

Therefore, we introduce BSEFs that compute the arguments' base scores given a preference ordering.

Definition 3. *A BSEF ν is a function that maps a set of preference orderings $\mathcal{P}$ to a set of base score functions $\mathcal{T}$ (for arguments in $\mathcal{X}$); namely $\nu : \mathcal{P} \to \mathcal{T}$.*

Obtaining the base scores of a set of arguments allows the transformation of BAFs into QBAFs.

3.4 Desirable Properties of Base Score Extraction Functions

In the following, we present the different axioms and properties of the BSEFs. First, we define the *Preference Coherence* axiom, which states that if one argument is preferred over another, the former's base score will be greater.

Axiom 1. (Preference Coherence) *Given a preference ordering $\succeq$, a BSEF $\nu(\succeq) = \tau$ satisfies* Preference Coherence *iff, for any arguments $a, b \in \mathcal{X}$:*

- *if $a \succ b$ or $a \succ\!\!\succ b$ then $\tau(a) > \tau(b)$; and*
- *if $a \simeq b$ then $\tau(a) = \tau(b)$.*

We now define *Preference Relation Coherence*, which provides a differentiation between the preference relations presented. Intuitively, if an argument is much preferred ($\succ\!\!\succ$) over another, the difference of their base scores will be greater than if the preference relation is only preferred ($\succ$).

Axiom 2. (Preference Relation Coherence) *Given a set of arguments arguments $a, b, c, d \in \mathcal{X}$ and the preference ordering $a \succ\!\!\succ b \succ c \simeq d$, a BSEF $\nu(\succeq) = \tau$ satisfies* Preference Relation Coherence *iff $\tau(a) - \tau(b) > \tau(b) - \tau(c) > \tau(c) - \tau(d)$.*

In our approach, the focus on extracting the base scores is set on the preference structure rather than the comparison between elements. The following axiom is necessary to define whether two or more preference orderings are isomorphic. If the preferences of two sets of arguments exhibit the same relationships, they can be said to have isomorphic preferences.

Definition 4. *Let $\mathcal{X}'$ be a set of arguments and $\succeq$ and $\succeq'$ be two preference orderings over $\mathcal{X}$ and $\mathcal{X}'$, respectively. Those orderings are isomorphic, denoted $\succeq \simeq \succeq'$, iff there is a bijective function $f : \mathcal{X} \to \mathcal{X}'$ such that $\forall i, j \in \mathcal{X},\ i \succeq j \Leftrightarrow f(i) \succeq' f(j)$.*

We require that a BSEF preserves the structural distinction of users' preferences. If two users have different (non-isomorphic) preference orderings, the assigned base scores must reflect it.

Axiom 3. (Preference Structure Coherence) *Given two preference orderings $\succeq$ and $\succeq'$ (over the same $\mathcal{X}$) that are not isomorphic ($\succeq \not\simeq \succeq'$),*

then a BSEF ν satisfies Preference Structure Coherence *iff* $\nu(\succeq) \neq \nu(\succeq')$.

Note that if two sets of arguments have different base scores, they will not necessarily have different preference ordering.

3.5 Design Choices for Base Score Extraction Functions

The presented axioms offer the necessary conditions that the BSEFs must satisfy, but determining the base scores is still a flexible and context-dependent challenge. This section outlines different design choices that must be considered when designing a BSEF.

3.5.1 Setting the Range. A choice that must be made is considering where the base scores of the most and least preferred arguments should be placed. Establishing them as 1 and 0, respectively, determines a strict and high importance of the preferences, but also allows for possible saturations from incoming supports or attacks, which would provoke information loss, an existing concern in PAFs. For a further investigation, we define the base score limits.

Definition 5. *The base score of the most preferred arguments from the BSEF is defined as* $\nu(\succeq)(x) = \top$ *for all* $x \in max_{\succeq}(\mathcal{X})$.[1]

Definition 6. *The base score of the least preferred arguments from the BSEF is defined as* $\nu(\succeq)(x) = \bot$ *for all* $x \in min_{\succeq}(\mathcal{X})$.[2]

The range is constrained to that of the gradual semantics, i.e. $[0, 1]$. Namely, it must be satisfied that $0 \leq \bot \leq \top \leq 1$. A BSEF will be considered normalised if its limits are set to $\top = 1$ and $\bot = 0$:

Property 1. (Base Score Normalisation) *A BSEF ν satisfies* Base Score Normalisation *iff* $\top = 1$ *and* $\bot = 0$.

In some scenarios, the limit base scores might be centred at 0.5 to differentiate between important and less important arguments.

Property 2. (Base Score Centralisation) *A BSEF ν satisfies* Base Score Centralisation *iff* $\top = 1 - \bot$.

The range influences the base scores' compression. For example, if the range is too low, the base scores would be very similar. Hence, the preferences of the arguments would not be properly reflected. The following example illustrates this concern.

Example 4. Given the arguments $a, b, c, d, e, f \in \mathcal{X}$, consider the preference ordering: $c \simeq f \succ b \simeq e \succ a \simeq d$. If the limits were set to $\top = 0.9$ and $\bot = 0.85$, the base scores could be placed only between $[0.85, 0.9]$ (e.g., $\tau(c) = \tau(f) = 0.9, \tau(b) = \tau(e) = 0.875, \tau(a) = \tau(d) = 0.85$). Although these base scores satisfy all the axioms, they may not sufficiently differentiate the preferences enough for some applications, since the difference between them is too low.

3.5.2 Preference Regularity. In many real-world cases, users can express their preferences regularly, meaning that the perceived difference in importance between two adjacent preferred arguments is roughly uniform across the ordering. For instance, if a user slightly prefers argument a over b, and b over c, the gap between a and b is comparable to that between b and c. Assuming such regularity simplifies the computation of base scores. Before introducing the property, we define the notion of adjacent arguments.

[1] $max_{\succeq}(\mathcal{X}) = \{x \in \mathcal{X} \mid \nexists y \in \mathcal{X} \; s.t. \; y \succ x \; or \; y \succ\!\!\succ x\}$.
[2] $min_{\succeq}(\mathcal{X}) = \{x \in \mathcal{X} \mid \nexists y \in \mathcal{X} \; s.t. \; x \succ y \; or \; x \succ\!\!\succ y\}$.

Definition 7. *Let $a, b \in \mathcal{X}$ and $\succeq$ be a preference ordering. The arguments a, b are adjacent in $\succeq$ iff one of the following conditions holds:*

- $a \succ b$ *and* $\nexists c$ *such that* $a \succ c \succ b$;
- $a \succ\!\!\succ b$ *and* $\nexists c$ *such that* $a \succ\!\!\succ c \succ b$ *or* $a \succ c \succ\!\!\succ b$ *or* $a \succ c \succ\!\!\succ b$;
- $a \simeq b$.

Property 3. (Base Score Relation Regularity) *A BSEF ν satisfies* Base Score Relation Regularity *iff for any preference relation $\succeq$ and any two pairs (a, b) and (c, d) of adjacent arguments in $\succeq$, then for $\tau = \nu(\succeq)$ it holds that $\tau(a) - \tau(b) = \tau(c) - \tau(d)$.*

Example 5. (Example 1 Cont.) Given the preference ordering: $c \simeq f \succ b \simeq e \succ a \simeq d$, let the limits be $\top = 0.9$ and $\bot = 0.1$. If the BSEF ν satisfies Base Score Relation Regularity, then any base score function obtained from ν would give base scores $\tau(c) = \tau(f) = 0.9, \tau(b) = \tau(e) = 0.5, \tau(a) = \tau(d) = 0.1$. Otherwise, the base scores $\tau(b)$ and $\tau(c)$ could be arbitrarily placed between 0.1 and 0.9.

If this regularity condition is not enforced, preference intensity could be encoded via irregular spacing between arguments rather than using a separate *much greater* relation. Yet, this flexibility makes the base score extraction process more difficult, as there is no clear rule for assigning consistent numerical differences.

Another property is that adding new arguments equally preferred to existing ones does not change the base scores of the other arguments.

Property 4. (Base Score Preference Stability) *Let $\succeq$ be a preference ordering, S a set of arguments such that $S \cap \mathcal{X} = \emptyset$, $\mathcal{X}' = \mathcal{X} \cup S$, and $\succeq'$ be a preference ordering over $\mathcal{X}'$ such that $\forall s \in S$, $\exists a \in \mathcal{X}$ such that $a \simeq s$ and, $\forall a, b \in \mathcal{X}$, $a \succeq' b$ iff $a \succeq b$. Then BSEFs ν (as per Definition 3) and ν' (mapping a set of preference orderings over $\mathcal{X}'$ to a set of base scores for arguments in $\mathcal{X}'$) satisfy* Base Score Preference Stability *iff for $\tau = \nu(\succeq)$ and $\tau' = \nu'(\succeq')$, for all $x \in \mathcal{X}$, it holds that $\tau(x) = \tau'(x)$.*

3.5.3 Preferences ratios. Users often express varying intensities in their preferences. For instance, one argument may be only slightly preferred over another, while another is much more important. To model such heterogeneous gaps, we introduce the concept of preference ratios, which quantify how much stronger a *much greater* preference should be compared to a regular preference ($\succ$). This ratio provides a systematic way to translate qualitative intensity (e.g., "much more important") into quantitative differences between base scores.

Example 6. (Example 5 Cont.) Now the preference ordering becomes: $c \simeq f \succ\!\!\succ b \simeq e \succ a \simeq d$. From asking the human, the ratio of the *much greater* preference is set to 3. The BSEF limits are set to 0.9 and 0.1. Then, the base scores could be set to $\tau(c) = \tau(f) = 0.9, \tau(b) = \tau(e) = 0.3, \tau(a) = \tau(d) = 0.1$. Here, the difference between preferred arguments is 0.2, while for much preferred arguments is 0.6, hence the ratio is respected.

Observe that the ratio between the different preferred relations and the limits must be placed carefully, as there is a risk of undermining the preferred or much more preferred relations. The following example shows two scenarios where this might occur:

Example 7. Consider $a, b, c, d \in \mathcal{X}$, with the preference ordering: $a \succ b \succ\succ c \succ d$. If the ratio is set to 1.33, with the limits at 0.75 and 0.25, the base scores would be $\tau(a) = 0.75, \tau(b) = 0.6, \tau(c) = 0.4, \tau(d) = 0.25$. The difference between preferred arguments is 0.15, while it is 0.2 for much more preferred arguments. At first sight, it does not seem representative enough. Alternatively, setting a ratio of 96 and the limits at 0.99 and 0.01, the base scores would be $\tau(a) = 0.99, \tau(b) = 0.98, \tau(c) = 0.02, \tau(d) = 0.01$. Here, the difference between the adjacent preferred arguments is 0.01, which intuitively may not be representative in many real-world contexts.

These examples highlight that the preference ratio serves as a tuning parameter controlling how sensitively the system reacts to *much greater* preferences. A well-chosen ratio helps capture meaningful differences in user priorities while maintaining numerical stability in base score extraction. The combination of these three different design choices thus gives a formal representation of how BSEFs may be personalised to different settings.

4 CONCRETE BASE SCORE EXTRACTION FUNCTIONS

In this section, two BSEFs are introduced. Those functions are based on the preference ordering of the arguments and allow for the presented design choices. An algorithm is developed to compute the base scores. Some examples of its use are shown using the running example. By definition, as the BSEFs are based on preference orderings, then, scenarios with only one argument are not considered.

Since preference orderings are monotonic, considering a descending preference ordering (from the most preferred to the least preferred), the BSEF should also be monotonically decreasing.

Definition 8. *A BSEF v is monotonically decreasing (increasing) iff for any base score function $\tau = v(\succeq)$, given two arguments a, b where $a \succeq b$, then $\tau(a) \leq \tau(b)$ ($\tau(a) \geq \tau(b)$, respectively).*

Proposition 1. *A monotonically increasing BSEF violates Axiom 1 for a descending preference ordering.*[3]

Creating a monotonically decreasing function directly from the preference ordering is possible. The following method consists of assigning a distance $d(x)$ to each argument $x \in \mathcal{X}$ according to the preference ordering and then normalising it between the desired range. The distance $d(x)$ is relative to the most preferred arguments, and the least preferred arguments are at the maximum absolute distance D to the most preferred arguments.

Base Score Extraction Function 1. (Adaptable range) Our first BSEF allows for the range to be set and is defined as follows. For any ordering $\succeq$ and $x \in \mathcal{X}$, let $\top$ be the base score for the most preferred arguments (see Definition 5) and $\bot$ be the base score of the least preferred arguments (see Definition 6). Then:

$$v_1(\succeq)(x) = \top + (\bot - \top) \cdot \frac{D - d(x)}{D - 1} \tag{1}$$

With the restriction $0 \leq \bot \leq \top \leq 1$ and $D > 1$, the base score functions obtained with this BSEF are well-defined, giving base scores inside $[0, 1]$. The edge cases occur at $d(x) = 1$ (for the most preferred arguments) and at $d(x) = D$ (for the least preferred arguments), yielding $v_1(\succeq)(x) = \top$ and $v_1(\succeq)(x) = \bot$ respectively.

The next BSEF consists of setting the compression of the base scores and the distances towards the edge cases.

Base Score Extraction Function 2. (Adaptable squeezing and distancing) Our next BSEF allows a parameter to determine how compressed the base scores are to be set, and another to determine how distant the edges are, and is defined as:

$$v_2(\succeq)(x) = \frac{D - d(x) + \alpha}{D - 1 + \beta} \tag{2}$$

where $\beta \in [0, +\infty)$ controls the compression of the base scores, and $\alpha \in [0, +\infty)$ determines the distance to the edges. Intuitively, larger values of β yield more compressed base scores, and larger values of α place the most preferred argument closer to 1. Note that it is necessary that $\alpha \leq \beta$, otherwise the base scores would be greater than 1. With $\beta = D - 1$, the spread of the base scores is compressed in a range of 0.5. This BSEF is better suited to cases where the number of arguments is unknown or likely to change, whereas the first direct control of the range and the edges allows for the range to be set directly, which provides more control over the edge cases.

Algorithm 1 shows how to extract the base scores from a set of arguments $\mathcal{X}$ and its preference ordering $\succeq$. We assume that there is at least one pair of arguments related by $\succ$ or $\succ\succ$ (i.e. $n+m > 0$, for n and m at lines 2 and 3 respectively), otherwise, since no preferences over the arguments exist, the base scores cannot be dictated from them. First, the algorithm assigns a distance $d = 1$ to the most preferred arguments. For each subsequent argument in the ordering, the assigned distance depends on the preference relation with its predecessor: if the arguments are equally preferred ($\simeq$), the distance remains unchanged. If the relationship between arguments is of strict preference ($\succ$), the distance is increased by δ, and if the relation is of much preferred ($\succ\succ$), the distance is increased by Δ.

Definition 9. *The distance between preferred arguments is $\delta \in (0, \infty)$, and for much more preferred arguments is $\Delta \in (0, \infty)$.*

After assigning this distance, the base scores are obtained using one of the proposed BSEFs. An example applying these distances in Alg. 1 is shown:

Example 8. The robot feeds a user with the following preference ordering: $c \simeq f \succ\succ b \simeq e \succ a \simeq d$. It confirms that the ratio for the *much greater* preference is 3. Then, the function parameters are set to $\delta = 1, \Delta = 3$. By choice, the limits are set to $\top = 0.8, \bot = 0.2$, and Property 3 is satisfied. The distances are $d(c) = d(f) = 1$, $d(b) = d(e) = 4$, and $d(a) = d(d) = 5$.[4] Since the limits are arbitrarily set, the BSEF to be used is v_1, and the base scores result in: $\tau(c) = \tau(f) = 0.8, \tau(b) = \tau(e) = 0.35, \tau(a) = \tau(d) = 0.2$.

The base scores obtained can be used for decision-making, as shown in the following example.

Example 9. (Example 8 Cont.) The final strengths for the decision arguments are computed based on the previous example base scores, using the QE Model semantics, which results in: $\sigma(D_1) = 0.54$ and $\sigma(D_2) = 0.40$. Then, the robot decides to move slowly. Note that

[3] The proof of this proposition is found in [16].

[4] $d(d)$ refers to the distance of argument d.

Algorithm 1 Flexible Normalised Base Score Extraction

Input: A set of arguments $\mathcal{X}$, its preference ordering $\succeq$, the increase for greater preference relation δ, for much greater preference relation Δ, the BSEF v_i (one of v_1 or v_2).
Output: Set of base scores $\mathcal{T}$ for the arguments in $\mathcal{X}$

$\mathcal{T} \leftarrow \{\}$ ▷ create an empty set of base scores
$n \leftarrow$ number of greater preferences ($\succ$) in $\succeq$
$m \leftarrow$ number of much greater preferences ($\succ\succ$) in $\succeq$
$D \leftarrow 1 + n \cdot \delta + m \cdot \Delta$
$d \leftarrow 1$ ▷ set the first distance value
for each argument $a \in \mathcal{X}$ ordered by $\succeq$ **do**
  if a is not most preferred **then**
    $b \leftarrow$ the previous argument in $\succeq$
    if $b \succ a$ **then**
      $d(a) \leftarrow d(a) + \delta$
    else if $b \succ\succ a$ **then**
      $d(a) \leftarrow d(a) + \Delta$
    else
      $d(a) \leftarrow d(a)$
  $\tau(a) \leftarrow v_i(\succeq)(a)$ ▷ Apply Eq. 1 for $i = 1$ or Eq. 2 for $i = 2$
  $\mathcal{T} \leftarrow \mathcal{T} \cup \{\tau(a)\}$
return $\mathcal{T}$

the user prefers arguments c and f, which increase the strength of the slow option and decrease the strength of the fast option, respectively. The arguments b and e increase and decrease the strength of the slow option, respectively. Finally, the arguments a and d aim to make the robot go fast. The algorithm's selection of moving the robot slowly accurately adapts to the user's preferences.

These BSEFs offer different benefits. First, both are flexible, allowing the base scores to be set in a desired range and proportionally adjusting the difference of base scores between arguments. Additionally, the parameters are easily modifiable, allowing corrections and learning from human feedback. Other functions could be used, e.g., an exponentially increasing or decreasing function, but there is a high chance that they do not adjust properly to a user's preferences. For example, given $a, b, c, d, e \in \mathcal{X}$ ordered as: $a \succ b \succ c \succ d \succ e$, an exponentially decreasing function could set the base scores at $\tau(a) = 1$, $\tau(b) = 0.9$, $\tau(c) = 0.75$, $\tau(d) = 0.5$, and $\tau(e) = 0.1$. Assuming that the distribution of the base scores follows that function is most probably incorrect. Allowing the user to confirm the ordering as: $a \succ b \succ c \succ d \succ\succ e$, the base scores could be set to $\tau(a) = 1$, $\tau(b) = 0.866$, $\tau(c) = 0.677$, $\tau(d) = 0.5$, $\tau(e) = 0.1$. These base scores are similar to the previous ones, but the confirmation of the large gap between arguments d and e provides a more accurate representation of the real preferences.

5 EVALUATION

This section presents a theoretical validation of the concrete BSEFs based on how they satisfy the theoretical properties we defined. We also undertake an experimental evaluation, showing the pros and cons of each BSEF. Finally, we give an intuition of which gradual semantics to select depending on the decision-making context.

v	A1	A2	A3	P1	P2	P3	P4
v_1	$\Delta > 0$ $\delta > 0$	$\Delta > \delta > 0$	✓	$\top = 1$ $\bot = 0$	$\top = 1 - \bot$	Δ and δ constants	✓
v_2	$\Delta > 0$ $\delta > 0$	$\Delta > \delta > 0$	✓	$\alpha = \beta = 0$	$\alpha = \beta/2$	Δ and δ constants	✓

Table 1: Summary of the conditions for satisfying the presented axioms (A) and desirable properties (P).

5.1 Theoretical Analysis

First, we show under which conditions our proposed BSEFs satisfy the axioms and properties presented. This is summarised in Table 1. The proofs can be found in [16].

For Axiom 1, constants Δ and δ must be greater than zero.

Proposition 2. *Setting the constants $\Delta > 0$ and $\delta > 0$ satisfies Axiom 1 for both BSEFs.*

For Axiom 2, it is necessary that the distance between much more preferred arguments is bigger than the distance between preferred arguments, and must be greater than zero.

Proposition 3. *Setting $\Delta > \delta > 0$ satisfies Axiom 2 for both BSEFs.*

Since the BSEFs are based on the distance between arguments given an ordering, if that ordering is changed, the base scores' ordering will also change.

Proposition 4. *Both BSEFs satisfy Axiom 3.*

Next, we focus on the BSEFs' desirable properties. First, on the range properties. First, in v_1.

Proposition 5. *Setting $\top = 1$ and $\bot = 0$ in v_1 satisfies* Base Score Normalisation *(Property 1).*

Proposition 6. *Setting $\top = 1 - \bot$ (with $\bot \leq 0.5$) in v_1 satisfies* Base Score Centralisation *(Property 2).*

Now, for the other BSEF v_2, those properties can be satisfied by adjusting the parameters α and β.

Proposition 7. *Setting $\alpha = \beta = 0$ in v_2 satisfies* Base Score Normalisation *(Property 1).*

Proposition 8. *Setting $\alpha = \frac{\beta}{2}$ in v_2 leads to equal limits $\top = 1 - \bot$, satisfying* Base Score Centralisation *(Property 2).*

To satisfy Property 3 (Base Score Relation Regularity), the algorithm variables to compute the distance (Δ, δ) must be constant.

Proposition 9. *Setting Δ and δ as fixed constants satisfies* Base Score Relation Regularity *(Property 3) for v_1 and v_2.*

Property 4 determines that the base scores of a set of arguments with a given preference relation remain constant when adding a new argument that is equally preferred to an existing one.

Proposition 10. *Both BSEFs satisfy* Base Score Preference Stability *(Property 4).*

Finally, if the selected gradual semantics is monotonic and balanced [10], if an argument is preferred to another and both have the same influences, the preferred argument will be stronger.

Preference ordering $\succeq$	Design Choices			QE Strengths		EB Strengths		DF Strengths	
	$\top$	$\bot$	Δ/δ	σ_{slow}	σ_{fast}	σ_{slow}	σ_{fast}	σ_{slow}	σ_{fast}
$a \simeq b \simeq c \simeq d \simeq e \simeq f$	-	-	-	0.5	**0.51**	0.50	**0.52**	0.44	**0.63**
$c \simeq f \succ\succ b \simeq e \succ a \simeq d$	0.9	0.1	3	**0.58**	0.33	**0.56**	0.39	**0.78**	0.23
	0.9	0.1	5	**0.57**	0.32	**0.55**	0.39	**0.79**	0.2
	0.75	0.25	5	**0.52**	0.4	**0.53**	0.43	**0.63**	0.38
$b \simeq e \succ a \simeq d \succ\succ c \simeq f$	0.8	0.2	3	0.5	**0.63**	0.52	**0.60**	0.28	**0.87**
	0.6	0.4	3	0.5	**0.53**	0.5	**0.54**	0.40	**0.71**
$a \simeq d \succ\succ c \simeq f \succ b \simeq e$	0.8	0.2	4	0.36	**0.6**	0.4	**0.59**	0.15	**0.76**
	1	0	4	0.25	**0.7**	0.37	**0.65**	0	**0.9**

Table 2: Examples of preference orderings for the running example with different design choices for the BSEF v_1 (Eq. 1). The value in bold between σ_{slow} and σ_{fast} is the system output. The (-) symbol indicates when a design choice is not needed, which is the case when all arguments are equally preferred, and their base scores are set to 0.5.

Proposition 11. *Given a BSEF v, where v satisfies Axiom 1, and a semantics σ, where σ satisfies monotonicity and balance, for any arguments a, b where $a \succ b$ and a and b have equivalent supporters and attackers, then $\sigma(a) > \sigma(b)$.*

5.2 Experimentation

The aim of this experiment is to evaluate the practical behaviour of the proposed BSEFs under different user preference profiles and design choices, and to examine how various gradual semantics influence the resulting decisions. Specifically, we test whether the generated QBAFs lead to consistent, interpretable, and preference-aligned outcomes across semantics.

We consider the running example from Fig. 1 in which the robot must choose between slow (D_1) and fast (D_2) feeding paces (see Example 1). Each QBAF instance is constructed using six arguments and a pair of decision arguments.

To explore the robustness of our approach, we generated 30,000 random samples of preference orderings and design choices, recording the option selected by three gradual semantics. These semantics are: QE Model, Euler-Based (EB) [3], and DF-QuAD (DF) [35], which are among the most used in the literature. The decision arguments' base scores are set to 0.5. We computed pairwise agreement and Cohen's Kappa coefficient (κ) to assess consistency between semantics. Under the centralisation property, the pairwise agreement values were $QE-EB = 0.98$, $QE-DF = 0.85$, and $EB-DF = 0.84$. The corresponding Cohen's Kappa scores were $QE-EB = 0.96$, $QE-DF = 0.65$, and $EB-DF = 0.62$. Without centralisation, the pairwise agreement values were $QE-EB = 0.96$, $QE-DF = 0.81$, and $EB-DF = 0.81$, with Cohen's Kappa scores of $QE-EB = 0.92$, $QE-DF = 0.52$, and $EB-DF = 0.52$. These results show a strong alignment between QE and EB, while DF-QuAD diverges more, particularly when base score ranges are extreme (e.g., $\tau(e) \approx 1$, $\tau(e) \approx 0$). This difference shows the structural properties of the methods discussed in Sec. 5.3.

To further analyse the influence of user preferences and design parameters, Table 2 reports examples of preference orderings combined with different design choices, evaluated under the three gradual semantics. The Base Score Relation Regularity property is assumed to reduce experimental complexity and noise.

The results confirm that the preference ordering itself is the dominant determinant of the final decision, while design choices act as fine-tuning factors that adjust numerical sensitivity but sometimes invert the selected option. For instance, when arguments related to safety (c and f) are prioritised over those concerning enjoyment (a and d), all semantics consistently select the slow feeding pace, aligning with the user's intended caution. Conversely, when enjoyment-related arguments are most preferred, the system reliably opts for fast feeding, demonstrating the coherence of the extracted base scores with user values.

Design choices such as the range limits affect the magnitude of the option strengths rather than the decision direction. Narrower ranges (e.g., $(\top, \bot) = (0.75, 0.25)$) compress the difference between argument strengths, producing less decisive but more stable outcomes, while broader ranges (e.g., $(1, 0)$) amplify contrasts and make the system more sensitive to preference changes. Similarly, the preference ratio (Δ/δ) controls how much stronger a *much greater* preference influences the base scores. Increasing this ratio enhances differentiation between preference tiers but may also overemphasise extreme preferences, especially under DF-QuAD, which is more responsive to high base score values.

5.3 Discussion on the Gradual Semantics

The previous analysis revealed that, although all semantics preserve preference–decision coherence, their sensitivity to base score variations differs substantially. This behaviour stems from the aggregation and influence functions each model employs.

Figure 2 shows the influence of a single attacker or supporter (the influencer), and an argument's base score (the influenced) for its final strength. The DF-QuAD semantics use product aggregation with a linear influence function, which causes the influencer's base score to dominate the final outcome. When an attacking (or supporting) argument has a base score close to 1, it almost completely drives the influenced argument's strength toward 0 (or 1), regardless of the influenced base score. This strong reactivity can be desirable in safety-critical scenarios, where highly preferred arguments should decisively determine the outcome, but it may reduce robustness in settings with uncertain or noisy preferences. In contrast, the QE semantics combines sum aggregation with a 2-Max influence function. Here, both the influencer and the influenced arguments contribute comparably to the final strength, leading to smoother transitions and more gradual adaptation to preference changes (e.g., given an influencer with a base score of 1, the influenced's final strength will be

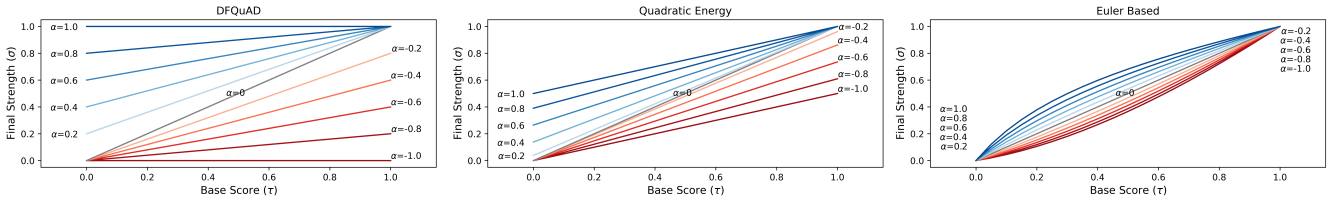

Figure 2: Influence of an argument's base score and its supporters or attackers on the final argument strength under different gradual semantics. Each subplot corresponds to a distinct semantics: DF-QuAD (left), Quadratic Energy (centre), and the Euler-based Semantics (right). The x-axis represents the base score of the influenced argument, and the y-axis its resulting final strength. Blue lines denote when the influencing argument is a supporter (+), while red lines denote when it is an attacker (-). Each line has assigned its aggregation (α), with colour intensity indicating its strength.

$\sigma = 0.5\tau + 0.5$ if it receives a support, and $\sigma = 0.5\tau$ if it receives an attack). QE, therefore, offers a good compromise between responsiveness and stability, which makes it particularly suitable for explainable decision-support systems that require consistent reasoning traces. Finally, the EB semantics is the most conservative, using sum aggregation with an Euler-based influence function. When an argument's base score is close to the extremes (0 or 1), its final strength remains nearly fixed at those values, regardless of new supports or attacks. This behaviour preserves high-confidence arguments, making EB appropriate when users insist on arguments being non-negotiable or when decisions must respect rigid value priorities.

6 RELATED WORK

Gradual argumentation relies on the arguments' base scores, and its literature has unveiled different ways of determining them depending on the context. In [11], the QuAD framework was introduced to support debates on design alternatives. Since these scenarios do not involve a large number of voters, base scores are assessed by experts or derived from predefined criteria. As an extension, in [34], the authors presented QuAD-V, which allows users to vote for or against arguments, using these votes to compute base scores. A similar procedure is carried out in [18]. In another work [17] in which a QBAF is generated from movie reviews, the base scores are obtained from an aggregation of critics' votes. Another use of QBAFs is to apply their methods for explainable AI, e.g. by transforming (deep) neural networks into QBAFs [1, 9]. For instance, in [1], the base scores of the arguments are obtained from the attacking, supporting neurons, and their activations in the network, similarly to [30]. In contrast, the work from [12] sets base scores according to users' preferences to personalise decision support, similarly to our work. There, base scores are initially set arbitrarily and then discounted for less preferred arguments. This work similarly relies on preferences, but uses them to determine the initial base scores directly.

Within structured argumentation, the work in [40] focuses on ASPIC argumentation frameworks with fuzzy set theory, using expert knowledge to assess argument importance, similarly to a base score.

In summary, two main limitations are found in existing work. First, most approaches do not account for the preferences of a single user when setting base scores, limiting personalisation in argumentation-based decision-making systems. Second, when preferences are considered, base scores are initialised arbitrarily and adjusted afterwards. In this work, we address those research gaps.

7 CONCLUSIONS AND FUTURE WORK

This work introduces BSEFs, a principled method to derive argument base scores from user preference orderings in gradual argumentation. We defined a set of axioms and desirable properties, proposed two concrete BSEFs with tunable design choices, and incorporated non-linear preference intensity to better capture human reasoning. Theoretical analysis confirmed that the functions satisfy coherence and regularity properties, while experiments in a robotics scenario showed that preference ordering dominated the decision outcome and design choices primarily modulate sensitivity.

Future directions include the fine-tuning or definition of the different design choices given the context and user feedback, allowing partial orderings in preferences in case some arguments cannot be compared [41], and computing the aggregation when using this approach when considering more than one user's preferences in the decision-making [31]. A philosophical question which should be commented on is whether it is realistic to consider the scenario in which the most preferred argument supports the least preferred argument. Our focus was on bipolar argumentation, but gradual structured argumentation frameworks are receiving attention lately [36, 37]. A future extension of this work is to adapt our method to structured argumentation. Furthermore, we plan to validate our approach by performing a user study. Finally, the extension of this work into value-based argumentation could provide an approach for more value-aligned decision-making [8, 13]. This could be potentially achieved by ordering the values promoted by the arguments and assigning a base score to each value.

ACKNOWLEDGMENTS

This work was supported by the project CHLOE-MAP PID2023-152259OB-I00 funded by MCIU/ AEI /10.13039/501100011033 and by ERDF, UE; the project ROBOCAT SDC006/25/000016 funded by the Generalitat de Catalunya, NextGenerationEU. A. Civit has been supported by AGAUR-FI ajuts (2023 FI-3 00065) Joan Oró of the Generalitat of Catalonia and the European Social Plus Fund. F. Toni has been supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934).

REFERENCES

[1] Emanuele Albini, Piyawat Lertvittayakumjorn, Antonio Rago, and Francesca Toni. 2020. Deep argumentative explanations. *arXiv preprint arXiv:2012.05766* (2020).

[2] Leila Amgoud. 2009. Argumentation for decision making. In *Argumentation in artificial intelligence*. Springer, 301–320.

[3] Leila Amgoud and Jonathan Ben-Naim. 2018. Evaluation of arguments in weighted bipolar graphs. *International Journal of Approximate Reasoning* (2018), 39–55.

[4] Leila Amgoud, Jonathan Ben-Naim, Dragan Doder, and Srdjan Vesic. 2016. Ranking arguments with compensation-based semantics. In *15th International Conference on Principles of Knowledge Representation and Reasoning (KR)*. 12–21.

[5] Leila Amgoud and Claudette Cayrol. 1998. On the acceptability of arguments in preference-based argumentation. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. 1–7.

[6] Leila Amgoud and Henri Prade. 2009. Using arguments for making and explaining decisions. *Artificial Intelligence* (2009), 413–436.

[7] Leila Amgoud and Srdjan Vesic. 2014. Rich preference-based argumentation frameworks. *International Journal of Approximate Reasoning* (2014), 585–606.

[8] Katie Atkinson and Trevor Bench-Capon. 2021. Value-based argumentation. *Journal of Applied Logics* (2021), 1543–1588.

[9] Hamed Ayoobi, Nico Potyka, and Francesca Toni. 2023. SpArX: Sparse Argumentative Explanations for Neural Networks. In *ECAI - 26th European Conference on Artificial Intelligence*. 149–156.

[10] Pietro Baroni, Antonio Rago, and Francesca Toni. 2019. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *International Journal of Approximate Reasoning* (2019), 252–286.

[11] Pietro Baroni, Marco Romano, Francesca Toni, Marco Aurisicchio, and Giorgio Bertanza. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation* (2015), 24–49.

[12] Elisa Battaglia, Pietro Baroni, Antonio Rago, and Francesca Toni. 2024. Integrating user preferences into gradual bipolar argumentation for personalised decision support. In *International Conference on Scalable Uncertainty Management*. Springer, 14–28.

[13] Gustavo A Bodanza and Esteban Freidin. 2023. Confronting value-based argumentation frameworks with people's assessment of argument strength. *Argument & Computation* (2023), 247–273.

[14] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*. Springer, 378–389.

[15] Aniol Civit, Antonio Andriella, Carles Sierra, and Guillem Alenyà. 2025. Multi-User Personalisation in Human-Robot Interaction: Resolving Preference Conflicts Using Gradual Argumentation. *arXiv preprint arXiv:2511.03576* (2025).

[16] Aniol Civit, Antonio Rago, Antonio Andriella, Guillem Alenyà, and Francesca Toni. 2026. From User Preferences to Base Score Extraction Functions in Gradual Argumentation (with Appendix). *arXiv preprint arXiv:2602.14674* (2026).

[17] Oana Cocarascu, Antonio Rago, and Francesca Toni. 2019. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. Association for Computing Machinery, 1261–1269.

[18] Louise Dupuis De Tarlé, Elise Bonzon, and Nicolas Maudet. 2022. Multiagent dynamics of gradual argumentation semantics. In *21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

[19] Jérôme Delobelle and Serena Villata. 2019. Interpretability of Gradual Semantics in Abstract Argumentation. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 15th European Conference, ECSQARU, Proceedings*. 27–38.

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

[21] Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2025. Argumentative Large Language Models for Explainable and Contestable Claim Verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 14930–14939.

[22] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* (2020), 411–437.

[23] Souhila Kaci, Leendert van Der Torre, and Serena Villata. 2018. Preference in abstract argumentation. In *Computational models of argument*. IOS Press, 405–412.

[24] Timotheus Kampik, Kristijonas Čyras, and José Ruiz Alarcón. 2024. Change in quantitative bipolar argumentation: sufficient, necessary, and counterfactual explanations. *International Journal of Approximate Reasoning* (2024), 109066.

[25] Francesco Leofante, Hamed Ayoobi, Adam Dejl, Gabriel Freedman, Deniz Gorur, Junqi Jiang, Guilherme Paulino-Passos, Antonio Rago, Anna Rapberger, Fabrizio Russo, et al. 2024. Contestable AI Needs Computational Argumentation. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. 888–896.

[26] Jean-Guy Mailly and Julien Rossit. 2020. Argument, I choose you! preferences and ranking semantics in abstract argumentation. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. 647–651.

[27] Sanjay Modgil and Henry Prakken. 2013. A general account of argumentation with preferences. *Artificial Intelligence* (2013), 361–397.

[28] Gregory B Northcraft, Jared N Preston, Margaret A Neale, Peter H Kim, and Melissa C Thomas-Hunt. 1998. Non-linear preference functions and negotiated outcomes. *Organizational Behavior and Human Decision Processes* (1998), 54–75.

[29] Nico Potyka. 2018. Continuous Dynamical Systems for Weighted Bipolar Argumentation. *KR* (2018), 148–57.

[30] Nico Potyka. 2021. Interpreting neural networks as quantitative argumentation frameworks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6463–6470.

[31] Antonio Rago, Oana Cocarascu, Joel Oksanen, and Francesca Toni. 2025. Argumentative review aggregation and dialogical explanations. *ARTIFICIAL INTELLIGENCE* (2025), 104291.

[32] Antonio Rago, Oana Cocarascu, and Francesca Toni. 2018. Argumentation-based recommendations: Fantastic explanations and how to find them. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. 1949–1955.

[33] Antonio Rago, Hengzhi Li, and Francesca Toni. 2023. Interactive Explanations by Conflict Resolution via Argumentative Exchanges. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. 582–592.

[34] Antonio Rago and Francesca Toni. 2017. Quantitative argumentation debates with votes for opinion polling. In *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 369–385.

[35] A Rago, F Toni, M Aurisicchio, and P Baroni. 2016. Discontinuity-free decision support with quantitative argumentation debates. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*. 63–72.

[36] Antonio Rago, Stylianos Loukas Vasileiou, Son Tran, Francesca Toni, and William Yeoh. 2025. A Methodology for Incompleteness-Tolerant and Modular Gradual Semantics for Argumentative Statement Graphs. In *Proceedings of the 22nd International Conference on Principles of Knowledge Representation and Reasoning*. 500–511.

[37] Anna Rapberger, Fabrizio Russo, Antonio Rago, and Francesca Toni. 2025. On Gradual Semantics for Assumption-Based Argumentation. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. 512–522.

[38] Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin Wainwright. 2014. When is it better to compare than to score? *arXiv preprint arXiv:1406.6618* (2014).

[39] Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir, and Antoni Perelló. 2019. Value alignment: a formal approach. *In Responsible Artificial Intelligence Agents Workshop (RAIA) in AAMAS* (2019).

[40] Nouredine Tamani and Madalina Croitoru. 2014. A quantitative preference-based structured argumentation system for decision support. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 1408–1415.

[41] Wietske Visser, Koen V Hindriks, and Catholijn M Jonker. 2011. An argumentation framework for qualitative multi-criteria preferences. In *International Workshop on Theorie and Applications of Formal Argumentation*. Springer, 85–98.

[42] Georgios N Yannakakis and John Hallam. 2011. Ranking vs. preference: a comparative study of self-reporting. In *International conference on affective computing and intelligent interaction*. Springer, 437–446.

[43] Xiang Yin, Nico Potyka, Antonio Rago, Timotheus Kampik, and Francesca Toni. 2025. Contestability in Quantitative Argumentation. *arXiv preprint arXiv:2507.11323* (2025).

[44] Xiang Yin, Nico Potyka, and Francesca Toni. 2024. CE-QArg: Counterfactual Explanations for Quantitative Bipolar Argumentation Frameworks. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. 697–707.