

ArguCast: A System for Online Multi-Forecasting with Gradual Argumentation

Deniz Gorur, Antonio Rago and Francesca Toni

Department of Computing, Imperial College London, UK

Abstract

Judgmental forecasting is a form of forecasting which employs (human) users to make predictions about specified future events. Judgmental forecasting has been shown to perform better than quantitative methods for forecasting, e.g. when historical data is unavailable or causal reasoning is needed. However, it has a number of limitations, arising from users' irrationality and cognitive biases. To mitigate against these phenomena, we leverage on computational argumentation, a field which excels in the representation and resolution of conflicting knowledge and human-like reasoning, and propose novel *ArguCast frameworks* (ACFs) and the novel online system *ArguCast*, integrating ACFs. ACFs and ArguCast accommodate *multi-forecasting*, by allowing multiple users to debate on multiple forecasting predictions simultaneously, each potentially admitting multiple outcomes. Finally, we propose a novel notion of *user rationality* in ACFs based on votes on arguments in ACFs, allowing the filtering out of irrational opinions before obtaining *group forecasting* predictions by means commonly used in judgmental forecasting.

Keywords

Bipolar Argumentation, Gradual Semantics, Judgmental Forecasting,

1. Introduction

Judgmental forecasting is a form of forecasting which employs (human) users to make predictions about specified future events [1]. It is advocated as a valuable alternative to conventional quantitative methods for forecasting when historical data is unavailable or causal reasoning is required [1]. However, judgmental forecasting has a number of limitations, arising from (human) users' irrationality and cognitive biases [2] arising from over-/under-confidence [3] in their judgment. To overcome these issues many solutions have been proposed. Researchers have investigated the best ways of eliciting probabilities from humans [4], how incentives and training change users' forecasting abilities [5], and the effect of scoring rules on users [1]. Another research direction has focused on employing many experts or humans and aggregating their predictions since it has been found that group judgment usually performs better as the impact of bias is reduced by cancelling random error [1]. However, when there are many humans involved in forecasting, a new problem arises: how to effectively combine all the predictions that are made. A further, orthogonal issue with existing systems, e.g. [6], is that any information provided by users, e.g. their forecasts or reasoning therefor, concern a single event, and thus must be provided separately for different events. It is easy to see that being able to consider


Arg&App 2023: International Workshop on Argumentation and Applications, September 2023, Rhodes, Greece

✉ d.gorur@imperial.ac.uk (D. Gorur); a.rago@imperial.ac.uk (A. Rago); ft@imperial.ac.uk (F. Toni)

🆔 0009-0008-8976-8919 (D. Gorur); 0000-0001-5323-7739 (A. Rago); 0000-0001-8194-1459 (F. Toni)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

different events in one forecasting framework would utilise the information provided by users much more effectively.

Meanwhile, *argumentation* constitutes a major component of human intelligence since the ability to engage in arguments is essential for humans to understand new problems, perform scientific reasoning, and express, clarify, and defend their opinions in their daily lives [7]. *Computational argumentation* (see [8, 9] for overviews) has become an important topic in artificial intelligence due to its ability to conjugate representational needs with user-related cognitive models and computational models for automated reasoning [10]. These computational models are formalised as *argumentation frameworks*. Argumentation involves reasoning with uncertainty, and resolving conflicting information so we posit that it is natural to apply techniques from argumentation to the area of judgmental forecasting. However, there has been very little research in the use of argumentation in forecasting technology in the past. We are aware of only one such approach [6], which restricts users' provided information to single events and single outcomes.

In order to address the aforementioned issues, and thus make contributions to the field of judgmental forecasting, we leverage on computational argumentation. Specifically, we propose the novel ArguCast frameworks (ACFs) and a novel online system ArguCast, integrating ACFs. Like [6], ACFs (and thus ArguCast) allows for groups of users to make forecasts on events while engaging in argumentative debates supported by votes on arguments exchanged in these debates, encouraging users to consider and share their reasoning for their forecasts. However, differently from the existing approach, ACFs accommodate multiple forecasting predictions with multiple outcomes from multiple users, allowing for information to be shared across events. We also propose a novel notion of *user rationality*, comprising vote rationality and prediction rationality in ACFs, allowing the filtering out of irrational opinions before obtaining *group forecasting* predictions. In doing so, we provide a novel, argumentative method for combining forecasts, which introduces *multi-forecasting* on multiple events simultaneously and accounts for human biases and rationality.

The structure of the paper is as follows. In Section 2, we describe the most relevant existing approaches to forecasting. In Section 3, we give the necessary background on computational argumentation, which is used for defining rationality. In Section 4, we define ACFs. In Section 5, we provide an overview of the implementation of ACFs as the ArguCast system. In Section 6, we define our notions of user rationality in ACFs and demonstrate how they can be used to filter out irrational predictions before we aggregate users' predictions. Finally, Section 7 concludes and considers possible future directions of work.

2. Related Work

In this section, we will discuss the most relevant of the existing approaches to forecasting from the literature.

There are two approaches to combining forecasts: the qualitative approach (e.g. a group discussion to reach consensus) and the mechanical approach (e.g. a simple or weighted average of the forecasts). It has been shown that those which are mechanical are more likely to lead to greater accuracy than those which are qualitative [2]. Many ways of mechanical combinations

methods have been proposed before such as linear, log-linear, and democratic opinion pools. The aggregation method that we use in Section 6 is a variation of the log-linear pooling method.

Some of the systems that attempt to improve the forecasting capability of users by incentives, e.g. via monetising the users’ predictions based on their accuracy, are Hypermind¹, Smarkets², PredictIt³, and Polymarket⁴. The idea behind all of these systems is based on prediction markets, where people bet on the predictions. Smarkets and PredictIt do not have any functionality for the agents to debate amongst each other, whereas Polymarket and Hypermind only have a general chat/forum.

Good Judgment Open (GJOpen)⁵ [11], Metaculus⁶, and Infer⁷ are examples of group judgment systems. They can all support binary and multiple-answer questions. In addition to these Metaculus supports numeric interval and date interval questions. They both have a comment section for users to put forward their reasoning for their forecast. GJOpen and Infer elicit both reasoning for why the forecast could be correct, and why the forecast could be wrong from users, as it has been shown that forcing users to think about why they might be wrong makes them better forecasters (see Appendix A of [12]). GJOpen also investigates what would happen if the top forecasters of their tournaments were put on teams called ‘Superforecasters’ [13], and they were found to outperform the simple average of the crowd.

However, all of the systems we have discussed lack any mechanisms for eliciting, representing and evaluating the argumentative reasoning which takes place in the debates amongst users. We are aware of only Irwin et al.[6] who have formalised an argumentation framework that supports forecasting. However, for all its strengths, this approach hosts a number of shortcomings, such as the fact that it can only handle questions with binary answers and does not allow the same argument to be used for multiple questions, or even in the same question. This could introduce repetition and sparsity, which could cause confusion in users. In our novel ArguCast frameworks, we address all of these issues.

3. Background

Our approach uses *Quantitative Bipolar Argumentation Frameworks (QBAFs)* [14]. A QBAF is a tuple $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau \rangle$ where $\langle \mathcal{X}, \mathcal{A}, \mathcal{S} \rangle$ is a *Bipolar Argumentation Framework (BAF)* [15] and τ is a *base score function*, such that:

- \mathcal{X} is a set of *arguments*;
- \mathcal{A} is a binary relation of *attack* on \mathcal{X} , $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$;
- \mathcal{S} is a binary relation of *support* on \mathcal{X} , $\mathcal{S} \subseteq \mathcal{X} \times \mathcal{X}$; and
- $\tau : \mathcal{X} \rightarrow [0, 1]$ is a total function; $\tau(a)$ is the *base score* of $a \in \mathcal{X}$.

¹<https://predict.hypermind.com>

²<https://smarkets.com/>

³<https://www.predictit.org/>

⁴<https://polymarket.com/>

⁵<https://www.gjopen.com>

⁶<https://www.metaculus.com>

⁷<https://www.infer-pub.com>

In this paper we focus on the *Discontinuity-Free QuAD gradual semantics (DF-QuAD)* [16] for QBAFs. DF-QuAD determines the strength of arguments based on combining their base scores and the aggregated strength of their attackers and supporters, where, for $a \in \mathcal{X}$, the *attackers of a* are $\mathcal{A}(a) = \{b \mid (b, a) \in \mathcal{A}\}$ and the *supporters of a* are $\mathcal{S}(a) = \{b \mid (b, a) \in \mathcal{S}\}$. Let the *strength aggregation function* be $\delta : [0, 1]^* \rightarrow [0, 1]$ such that, for $T = (v_1, \dots, v_n) \in [0, 1]^*$:

$$\begin{aligned} \text{if } n = 0 & : \delta(T) = 0; \\ \text{if } n = 1 & : \delta(T) = v_1; \\ \text{if } n = 2 & : \delta(T) = f(v_1, v_2); \\ \text{if } n > 2 & : \delta(T) = f(\delta(v_1, \dots, v_{n-1}), v_n) \end{aligned}$$

where, for $x, y \in [0, 1]$, $f(x, y) = x + (1 - x) \cdot y = x + y - x \cdot y$. Let the *combination function* be defined as $c : [0, 1] \times [0, 1] \times [0, 1] \rightarrow [0, 1]$, where for $v_0, v_a, v_s \in [0, 1]$:

$$\begin{aligned} c(v_0, v_a, v_s) &= v_0 - v_0 \cdot |v_s - v_a| && \text{if } v_a \geq v_s; \\ c(v_0, v_a, v_s) &= v_0 + (1 - v_0) \cdot |v_s - v_a| && \text{if } v_a < v_s. \end{aligned}$$

Then, DF-QuAD computes the strength of arguments by the *score function* $\sigma : \mathcal{X} \rightarrow [0, 1]$ where, for any $a \in \mathcal{X}$, $\sigma(a) = c(\tau(a), \delta(\sigma(\mathcal{A}(a))), \delta(\sigma(\mathcal{S}(a))))$ such that $\sigma(\mathcal{A}(a)) = (\sigma(a_1), \dots, \sigma(a_n))$, where (a_1, \dots, a_n) is an arbitrary permutation of the ($n \geq 0$) attackers in $\mathcal{A}(a)$, and $\sigma(\mathcal{S}(a)) = (\sigma(s_1), \dots, \sigma(s_m))$, where (s_1, \dots, s_m) is an arbitrary permutation of the ($m \geq 0$) supporters in $\mathcal{S}(a)$.

4. ArguCast Frameworks

We introduce novel ArguCast frameworks, accommodating multi-forecasting, i.e. multiple *forecasting predictions* with multiple outcomes from multiple *users*, supported by argumentative debates and *votes* on arguments exchanged in these debates.

Definition 1. An ArguCast framework (ACF) is a tuple $\langle \mathcal{X}, \mathcal{R}, \mathcal{U}, \mathcal{V}, \mathcal{P} \rangle$ such that:

- $\mathcal{X} = \mathcal{F} \cup \mathcal{D}$ is a finite set of arguments where \mathcal{F} and \mathcal{D} are disjoint; elements of \mathcal{F} and \mathcal{D} are referred to, respectively, as *forecasting and non-forecasting arguments*;
- $\mathcal{R} = \mathcal{A} \cup \mathcal{S} \subseteq \mathcal{D} \times \mathcal{X}$, where \mathcal{A} and \mathcal{S} are disjoint relations (i.e. sets of pairs from $\mathcal{D} \times \mathcal{X}$) of *attack and support, respectively*;
- \mathcal{U} is a finite set of users;
- $\mathcal{V} : \mathcal{U} \times \mathcal{D} \rightarrow \{-, +\}$ is a (partial) function; $\mathcal{V}(u, a)$ is the vote of user $u \in \mathcal{U}$ on (non-forecasting) argument $a \in \mathcal{D}$;
- $\mathcal{P} : \mathcal{U} \times \mathcal{F} \rightarrow [0, 1]$ is a (partial) function; $\mathcal{P}(u, b)$ is the forecasting prediction by user $u \in \mathcal{U}$ on (forecasting) argument $b \in \mathcal{F}$.

Forecasting arguments represent answers to forecasting questions. There may be any number of forecasting arguments, as the answers may be Yes/No or take any value in a discrete set (thus the forecasting predictions may have multiple outcomes). If there is a single forecasting question of

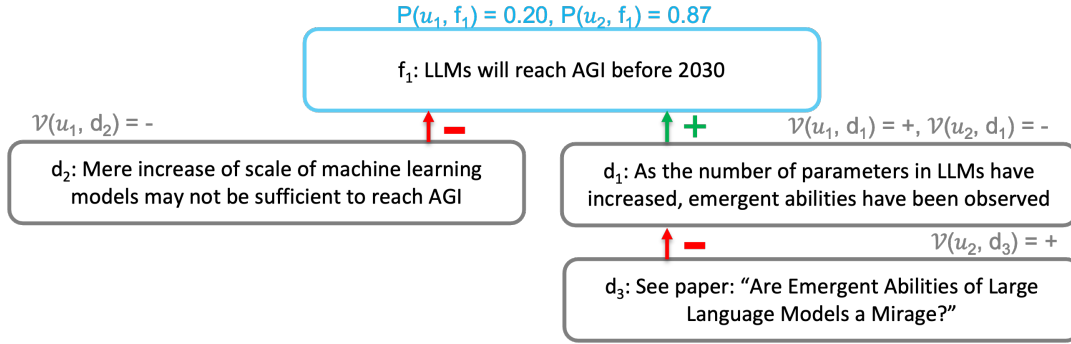


Figure 1: A visual representation of Example 1, where attacks and supports are represented by red and green, respectively, edges.

interest admitting a binary answer (e.g. ‘Will Miami Heat win the 2022-23 NBA Championship?’) then we assume that \mathcal{F} is a singleton set consisting of an argument for the positive answer to the question. If there are multiple forecasting questions or a single forecasting question admitting several alternative answers (e.g. ‘Which team will win the 2022-23 NBA Championship?’) then \mathcal{F} will consist of several forecasting arguments. Non-forecasting arguments can be seen as the users’ rationales/opinions around the forecasting arguments. Note that, by definition of \mathcal{A} and \mathcal{S} , forecasting arguments can be attacked/supported but they cannot attack/support other arguments, whereas non-forecasting arguments can attack/support (or be attacked/supported by) any arguments, including potentially attacking/supporting more than one argument. The users are forecasters, who can vote (positively or negatively) on non-forecasting arguments and/or express a numerical prediction (in $[0,1]$) for forecasting arguments. The votes indicate agreement or disagreement with the non-forecasting argument, whereas prediction forecasts indicate the users’ degree of belief in the forecasting arguments. Note that, as \mathcal{V} and \mathcal{P} may be partial, users may refrain from voting and forecasting.

Example 1. A possible ACF for the question ‘Will large language models (LLMs) reach AGI before 2030?’ is $\mathcal{F} = \{f_1 = \text{‘LLMs will reach AGI before 2030’}\}$, $\mathcal{D} = \{d_1 = \text{‘As the number of parameters in LLMs have increased, emergent abilities have been observed’}, d_2 = \text{‘Mere increase of scale of machine learning models may not be sufficient to reach AGI’}, d_3 = \text{‘See paper: “Are Emergent Abilities of Large Language Models a Mirage?”’}\}$, $\mathcal{A} = \{(d_2, f_1), (d_3, d_1)\}$, $\mathcal{S} = \{(d_1, f_1)\}$, $\mathcal{U} = \{u_1, u_2\}$, $\mathcal{V}(u_1, d_1) = +$, $\mathcal{V}(u_1, d_2) = -$, $\mathcal{V}(u_2, d_1) = -$, $\mathcal{V}(u_2, d_3) = +$, $\mathcal{P}(u_1, f_1) = 0.87$, $\mathcal{P}(u_2, f_1) = 0.20$, as shown in Figure 1.

Finally, note that our ACFs share some features of existing argumentation frameworks, but are different therefrom as follows. Like BAFs [15], ACFs use two relations of attack/support but in addition ACFs distinguish two types of arguments (forecasting and non-forecasting arguments) and include users, users’ votes on non-forecasting arguments and users’ predictions on forecasting arguments. Like QuAD frameworks [17], ACFs single out a specific type of argument under debate (forecasting arguments in ACFs but answer arguments in [17]) but

QuAD frameworks also distinguish con/pro arguments and admit a single relation whereas we distinguish attack/support relations, allowing for arguments to potentially attack some arguments and support some others. Also, QuAD frameworks lack users, votes and predictions, but include base scores for arguments, absent in ACFs (where, however, they can be obtained using votes, see below). QuAD-V [18] is an extension of QuAD frameworks like ACFs including users and votes and excluding base scores. While the votes in QuAD-V are given by a total function into $\{-, ?, +\}$, we use a partial function into $\{-, +\}$. QuAD-V frameworks also lack support for forecasting predictions. Like the *forecasting argumentation frameworks* (FAFs) of [6], ACFs are designed to support forecasting but FAFs can only handle questions with binary answers (as they can only have one proposal argument at a time). Like FAFs, ACFs single out a specific type of argument under debate (forecasting arguments in ACFs but *proposal arguments* in [6]), and they also support users with votes and forecasts. The votes in FAFs are assigned by a total function that forces users to provide their opinion on every argument. FAFs also use *amendment arguments* (arguments proposing the forecasted probability is increased or decreased) as well as pro/con arguments as in QuAD and V-QuAD frameworks, and a single relation between arguments, where amendment arguments can only relate to proposal arguments and con/pro arguments can only relate to amendment and other con/pro arguments. A further difference between ACFs and FAFs lies in the fact that, like in QuAD and QuAD-V, FAFs distinguish con/pro arguments and admit a single relation, which could introduce repetition and sparsity, which will lead to confusion in users. ACFs avoid this issue by adopting attack/support relations rather than a single relation type.

5. ArguCast

ArguCast is an online system, available at <https://argucast.herokuapp.com>, accommodating ACFs in practice as the basis for judgemental forecasting.⁸ We focus here on the system’s functionalities. Note that, even though the formalisation of ACFs handles binary and multi-answer forecasting questions, ArguCast supports only binary questions currently. Also, whereas ACFs allow for the same non-forecasting arguments to contribute to debating several forecasting arguments, ArguCast assumes for the time being that each non-forecasting argument contributes to debating only one forecasting argument. Thus, each ACF in ArguCast can be seen as the composition of disjoint ACFs, one for each forecasting question.

ArguCast is login-protected so all users need to register before engaging in forecasting. Users can add their own forecasting questions (with accompanying forecasting arguments, amounting to a positive answer to each question, as illustrated in Figure 2), or select from currently active questions (as illustrated in Figure 3, also showing that users can search for specific questions and add new questions by clicking on the plus button).

Figure 4 shows ArguCast’s representation of the ACF for one of the forecasting questions in Figure 3 from the viewpoint of a single user (i.e. $u \in \mathcal{U}$). Note that users do not have access to other users’ votes and predictions but they can see everything else. Note that ArguCast

⁸ArguCast’s user interface is implemented with React.js (<https://react.dev>). Storage of arguments, users, and predictions were on a PostgreSQL database. The Web API that connects to the database and executes queries requested from the user interface was implemented with Python’s Flask library.

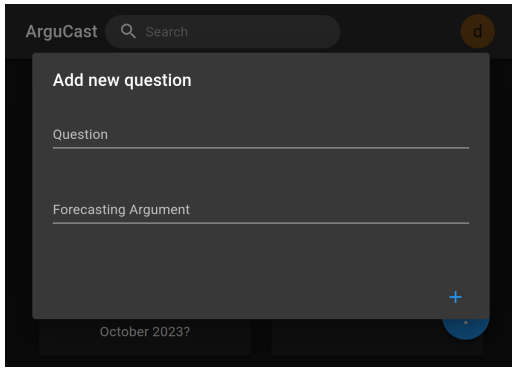


Figure 2: Dialogue to add new questions.

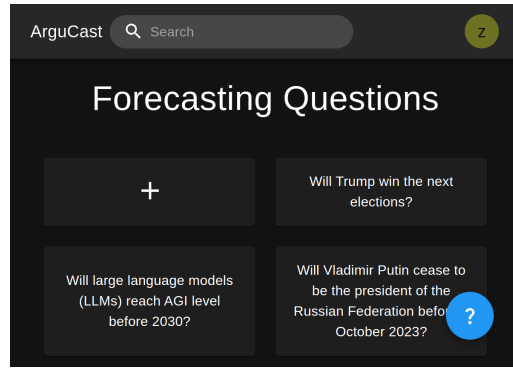


Figure 3: Overview of the active questions.

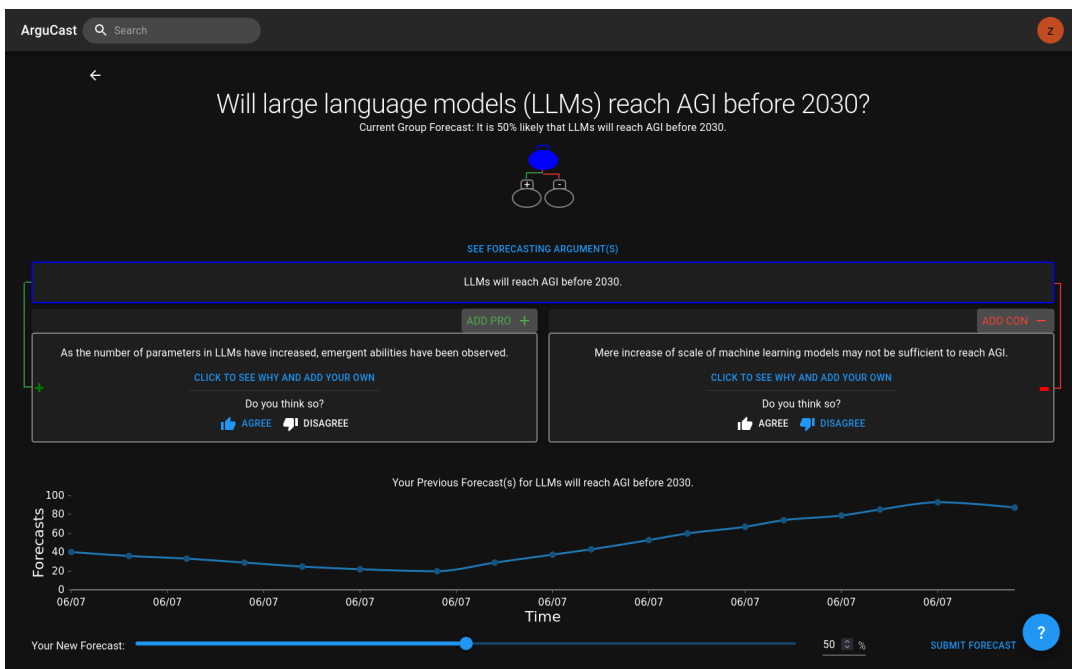


Figure 4: A fragment of Example 1 represented in ArguCast, from the viewpoint of user u_1 . Here we show arguments $f_1/d_1/d_2$, u_1 's votes on d_1/d_2 , and u_1 's forecasting predictions on f_1 over time.

supports two tree-based visualisations of each debate/ACF: a global, abstract visualisation (as shown at the top of Figure 4), focusing on the relations between the arguments in the ACF; and a localised visualisation around specific arguments (as shown in the centre of Figure 4), showing their attackers and supporters and allowing the user to vote (cf. \mathcal{V}). The localised visualisation supports the addition of supporting and attacking arguments. In both visualisations, forecasting arguments are outlined in blue and non-forecasting arguments are outlined in grey.

The attack/support relations (i.e. \mathcal{A}/\mathcal{S}) between arguments are represented as red/green edges, respectively, with a minus/plus (-/+) sign, respectively.

Below the question in Figure 4, the current group forecast is shown. The group forecast does not change in our system as of yet as the aggregation of users' predictions (which will be defined in Section 6) is not implemented in ArguCast.

Below the debate, the user has the ability to put forward their forecasting prediction (i.e. \mathcal{P}) for the forecasting argument using the slider. The slider ranges from $[0, 100]\%$, which maps to $[0, 1]$. The user can change their vote as they please, iteratively.

Any user can see the debates already present in the system. However, in order to add new arguments, cast opinions by voting, and put forward predictions, the user needs to be signed in to their account. If a user does not have an account they can sign up with their email or Google account.

6. Extensions of ArguCast Frameworks

The ArguCast frameworks are the base for improving forecasting systems using argumentation. One way of doing so is to define notions of rationality so we can filter out irrational users when aggregating forecasting predictions, which we will now demonstrate. Note that filtering for rationality and the aggregation of users' forecasts has not, as of yet, been implemented into ArguCast system.

In the remainder, unless specified otherwise, we will assume as given an ACF $\langle \mathcal{X}, \mathcal{R}, \mathcal{U}, \mathcal{V}, \mathcal{P} \rangle$. We will also assume that \mathcal{R} in this ACF is acyclic.

An ACF captures the opinions of all users (in \mathcal{U}) involved in forecasting. We can filter out the opinions of individual users as *user QBAFs*, i.e. a QBAF representing a single user's votes in the ACF, and then apply gradual semantics thereto for determining *rationality* of users by comparing the strengths of arguments and votes/forecasting predictions.

Definition 2. A user QBAF for $u \in \mathcal{U}$ is a QBAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau_u \rangle$ such that, for $a \in \mathcal{X}$:

$$\tau_u(a) = \begin{cases} 1 & \text{if } \mathcal{V}(u, a) = +, \\ 0 & \text{if } \mathcal{V}(u, a) = -, \\ 0.5 & \text{if } \mathcal{V}(u, a) \text{ is undefined.} \end{cases}$$

The attacking and supporting strengths of an argument $a \in \mathcal{X}$ in the user QBAF are defined as $\delta(\sigma(\mathcal{A}(a)))$ and $\delta(\sigma(\mathcal{S}(a)))$, denoted $\sigma_{\mathcal{A}}^u(a)$ and $\sigma_{\mathcal{S}}^u(a)$.

Example 2. A user QBAF for u_1 for the ACF given in Example 1 would be $\mathcal{X} = \{f_1, d_1, d_2, d_3\}$, $\mathcal{A} = \{(d_2, f_1), (d_3, d_1)\}$, $\mathcal{S} = \{(d_1, f_1)\}$, $\tau_{u_1}(f_1) = 0.5$, $\tau_{u_1}(d_1) = 1$, $\tau_{u_1}(d_2) = 0$, and $\tau_{u_1}(d_3) = 0.5$. Using DF-QuAD the strength of arguments is $\sigma(d_3) = 0.5$, $\sigma(d_2) = 0$, $\sigma(d_1) = c(1, 0.5, 0) = 1 - 1 \cdot |0 - 0.5| = 0.5$, and $\sigma(f_1) = c(0.5, 0, 0.5) = 0.5 + (1 - 0.5) \cdot |0.5 - 0| = 0.75$. Then attacking and supporting strengths of the non-forecasting argument d_1 is $\sigma_{\mathcal{A}}^u(d_1) = 0.5$ and $\sigma_{\mathcal{S}}^u(d_1) = 0$, respectively.

We define two notions of *user rationality* for ACFs: *vote rationality*, which compares the vote of a user on any non-forecasting argument with its strength; and *prediction rationality*, which compares the user's forecasting prediction on any forecasting argument with its strength. In the remainder, we will assume as given a user QBAF $\langle \mathcal{X}, \mathcal{A}, \mathcal{S}, \tau_u \rangle$ for a user $u \in \mathcal{U}$ in the ACF.

Definition 3. User u is vote rational iff $\forall a \in \mathcal{X}$:

$$\begin{aligned} \text{if } \mathcal{V}(u, a) = - \text{ then } \sigma_{\mathcal{A}}^u(a) &\geq \sigma_{\mathcal{S}}^u(a); \\ \text{if } \mathcal{V}(u, a) = + \text{ then } \sigma_{\mathcal{A}}^u(a) &\leq \sigma_{\mathcal{S}}^u(a). \end{aligned}$$

Example 3. Continuing Example 2, user u_1 is not vote rational. The user agreed with d_1 and the strength of the attacking arguments is bigger than the strength of the supporting arguments (i.e. $\sigma_{\mathcal{A}_{u_1}}(d_1) > \sigma_{\mathcal{S}_{u_1}}(d_1)$). In this instance, the vote rationality forces the user to add reasoning for the argument d_1 or put forward their opinion on d_3 .

Definition 4. User u is prediction rational iff $\forall a \in \mathcal{A}$:

$$\begin{aligned} \text{if } \sigma(a) < 0.5 \text{ then } \mathcal{P}(u, a) &< 0.5; \\ \text{if } \sigma(a) > 0.5 \text{ then } \mathcal{P}(u, a) &> 0.5; \\ \text{if } \sigma(a) = 0.5 \text{ then } \mathcal{P}(u, a) &= [0.5 - \epsilon, 0.5 + \epsilon] \text{ for some small } \epsilon. \end{aligned}$$

Example 4. Continuing Example 2, user u_1 is not prediction rational. User u_1 's forecasting prediction is $P(u_1, f_1) = 0.2$ and the strength of the forecasting argument is $\sigma(f_1) = 0.75$. In this instance, u_1 needs to change its forecasting prediction to be below 0.5 or update its vote(s) to decrease the strength of f_1 . This demonstrates how prediction rationality requires that a user's forecasting predictions be in line with their votes.

Definition 5. The ACFs are collectively rational iff $(\forall u \in \mathcal{U})$ are vote rational and prediction rational.

Aggregation of forecasts requires all the agents to be collectively rational. The process of aggregation thus uses only the forecasting predictions.

We use a weighted aggregation function where the weights are *Brier scores* [19] which represent the accuracy of each user in the previous questions that have an outcome. So, the users with better Brier scores will have a greater influence on the aggregated prediction. The outcome of each question is represented by $O_i \in \{0, 1\}$, where $O_i = 1$ if the outcome was true and $O_i = 0$ if the outcome was false.

Definition 6. Given all N outcomes as a set $(\{O_1, \dots, O_N\})$ and the corresponding forecasting predictions for user $u \in \mathcal{U}$ $(\{P(u)_1, \dots, P(u)_N\})$, the Brier score of u is:

$$b_u = \frac{1}{N} \sum_{t=1}^N (P_t - O_t)^2$$

Brier scores are the mean squared error of the user’s forecasting accuracy. A low b_u represents higher accuracy and a high b_u represents lower accuracy.

Then, our aggregation function is an adaptation of [20] where we also use (the negation of the) Brier scores to obtain a weighted aggregation.

Definition 7. The geometric mean of odds with systematic bias $\alpha \geq 1$, $\Omega : ACF \rightarrow [0, 1]$ is:

$$\Omega(ACF) = \left[\sqrt[|\mathcal{U}|]{\prod_{u \in \mathcal{U}} \left(e^{(1-b_u)} \frac{\mathcal{P}(u)}{1 - \mathcal{P}(u)} \right)} \right]^\alpha$$

The aggregation function $\lambda : ACF \rightarrow [0, 1]$ is:

$$\begin{aligned} \text{if } |\mathcal{U}| \neq 0 : & \quad \lambda(ACF) = \frac{\Omega(ACF)}{\Omega(ACF) + 1} \\ \text{otherwise} & \quad \lambda(ACF) = 0 \end{aligned}$$

The geometric mean of odds has been shown (empirically) to outperform [20] the arithmetic mean of odds as uncertain predictions will have less influence. Note also that if the systematic bias is 1 then the geometric mean of odds is similar to the arithmetic mean of odds. We will use $\alpha = 2.5$ for simplicity, however in practice, the value of α could be estimated [20].

7. Conclusions and Future Work

We have introduced our novel ACFs, accommodating forecasting predictions from users, argumentative debates (as Bipolar AFs) amongst users, and votes on arguments exchanged in these debates. We also described ArguCast, our online platform which instantiates ACFs. Then, we defined our notions of rational users for ACFs and showed how we can filter out irrational users when we combine users’ predictions. We have also shown a way to combine users’ predictions using the geometric mean of odds weighted by users’ Brier scores.

ACFs open up numerous avenues for future work. First, we plan to implement rationality constraints and prediction aggregation (in the forms discussed in Section 6 as well as others) in our online system and then empirically evaluate how much the accuracy of the forecasts improves, comparing with those defined in [6]. Second, we will build on the fact that ACFs provide the formal basis for further theoretical developments combining forecasting and argumentation. For example, at the moment, ACFs only allow users to vote on non-forecasting arguments so that we can apply rationality constraints to users. However, we would like to see how we can accommodate votes on attack/support relations to capture the users’ beliefs on the relations between arguments, which would allow us to extend the rationality constraints we have introduced. Another possible theoretical development would be to include a mapping from users to their contributed arguments to assess how this ownership affects voting, possibly allowing us to model users’ cognitive biases, such as *confirmation bias* [21], with argumentation, as in [22]. Finally, it would be interesting to generate explanations for the combined prediction, leveraging on argumentation’s amenability for explanation (see [23, 24] for recent surveys on its application to explainable AI).

Acknowledgments

This research was partially funded by the ERC under the EU's Horizon 2020 research and innovation programme (grant agreement no. 101020934, ADIX), by J.P. Morgan and by the Royal Academy of Engineering, UK, under the Research Chairs and Senior Research Fellowships scheme. Any views or opinions expressed herein are solely those of the authors.

References

- [1] M. Zellner, A. E. Abbas, D. V. Budescu, A. Galstyan, A survey of human judgement and quantitative forecasting methods, *Royal Society Open Science* 8 (2021) rsos.201187, 201187. doi:10.1098/rsos.201187.
- [2] M. Lawrence, P. Goodwin, M. O'Connor, D. Önkal, Judgmental forecasting: A review of progress over the last 25years, *International Journal of Forecasting* 22 (2006) 493–518. doi:10.1016/j.ijforecast.2006.03.007.
- [3] D. A. Moore, P. J. Healy, The trouble with overconfidence., *Psychological Review* 115 (2008) 502–517. doi:10.1037/0033-295X.115.2.502.
- [4] T. S. Wallsten, D. V. Budescu, A review of human linguistic probability processing: General principles and empirical evidence, *The Knowledge Engineering Review* 10 (1995) 43–62. doi:10.1017/S0269888900007256.
- [5] W. Chang, E. Chen, B. Mellers, P. Tetlock, Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments, *Judgment and Decision Making* 11 (2016) 509–526. doi:10.1017/S1930297500004599.
- [6] B. Irwin, A. Rago, F. Toni, Argumentative forecasting, in: P. Faliszewski, V. Mascardi, C. Pelachaud, M. E. Taylor (Eds.), *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022, 2022*, pp. 1636–1638. doi:10.5555/3535850.3536060.
- [7] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence* 77 (1995). doi:10.1016/0004-3702(94)00041-X.
- [8] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. R. Simari, M. Thimm, S. Villata, Towards artificial argumentation, *AI Magazine* 38 (2017) 25–36.
- [9] P. Baroni, D. Gabbay, M. Giacomin, L. van der Torre (Eds.), *Handbook of Formal Argumentation*, College Publications, 2018.
- [10] M. Lippi, P. Torroni, Argumentation mining: State of the art and emerging trends, *ACM Transactions on Internet Technology* 16 (2016). doi:10.1145/2850417.
- [11] P. E. Tetlock, B. A. Mellers, N. Rohrbaugh, E. Chen, Forecasting tournaments: Tools for increasing transparency and improving the quality of debate, *Current Directions in Psychological Science* 23 (2014). doi:10.1177/0963721414534257.
- [12] C. W. Karvetski, C. Meinel, D. T. Maxwell, Y. Lu, B. A. Mellers, P. E. Tetlock, What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters?, *International Journal of Forecasting* 38 (2022) 688–704. doi:https://doi.org/10.1016/j.ijforecast.2021.09.003.

- [13] P. E. Tetlock, D. Gardner, *Superforecasting: The art and science of prediction*, Random House, 2016.
- [14] P. Baroni, A. Rago, F. Toni, From fine-grained properties to broad principles for gradual argumentation: A principled spectrum, *International Journal of Approximate Reasoning* 105 (2019) 252–286. doi:10.1016/j.ijar.2018.11.019.
- [15] C. Cayrol, M. C. Lagasquie-Schiex, On the Acceptability of Arguments in Bipolar Argumentation Frameworks, in: D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, L. Godo (Eds.), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 3571, Berlin, Heidelberg, 2005, pp. 378–389. doi:10.1007/11518655_33, series Title: *Lecture Notes in Computer Science*.
- [16] A. Rago, F. Toni, M. Aurisicchio, P. Baroni, Discontinuity-free decision support with quantitative argumentation debates, in: *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning, KR’16*, AAAI Press, 2016, p. 63–72.
- [17] P. Baroni, M. Romano, F. Toni, M. Aurisicchio, G. Bertanza, Automatic evaluation of design alternatives with quantitative argumentation, *Argument & Computation* 6 (2015) 24–49. doi:10.1080/19462166.2014.1001791.
- [18] A. Rago, F. Toni, Quantitative Argumentation Debates with Votes for Opinion Polling, in: B. An, A. Bazzan, J. Leite, S. Villata, L. Van Der Torre (Eds.), *PRIMA 2017: Principles and Practice of Multi-Agent Systems*, volume 10621, Cham, 2017, pp. 369–385. doi:10.1007/978-3-319-69131-2_22, series Title: *Lecture Notes in Computer Science*.
- [19] G. W. BRIER, Verification of forecasts expressed in terms of probability, *Monthly Weather Review* 78 (1950) 1 – 3. doi:https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
- [20] V. A. Satopää, J. Baron, D. P. Foster, B. A. Mellers, P. E. Tetlock, L. H. Ungar, Combining multiple probability predictions using a simple logit model, *International Journal of Forecasting* 30 (2014) 344–356. doi:https://doi.org/10.1016/j.ijforecast.2013.09.009.
- [21] R. S. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises, *Review of General Psychology* 2 (1998) 175 – 220.
- [22] A. Rago, H. Li, F. Toni, Interactive explanations by conflict resolution via argumentative exchanges, *CoRR abs/2303.15022* (2023). doi:10.48550/arXiv.2303.15022. arXiv:2303.15022.
- [23] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, *The Knowledge Engineering Review* 36 (2021) e5. doi:10.1017/S0269888921000011.
- [24] K. Cyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: A survey, in: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 2021, pp. 4392–4399. doi:10.24963/ijcai.2021/600.