On Monotonic and Nonmonotonic Effects of Arguments in Public Interest Communication

Pietro Baroni[1,*], Giulio Fellin[1], Massimiliano Giacomin[1] and Carlo Proietti[2]

[1]*DII - University of Brescia, Italy*
[2]*ILC - Consiglio Nazionale delle Ricerche*

Abstract

In a previous work we introduced a vector-based extension of value-based argumentation for public interest communication aimed to provide an articulated model of the impact of a communication campaign on a set of target audiences. The proposed model was monotonic, intuitively meaning that adding arguments to a campaign and enlarging the set of the values they cover cannot decrease the effectiveness of the campaign itself. As this property does not always hold in practice, in this paper we extend the model in order to encompass nonmonotonic effects both at the level of quantitative measures of campaign impact and of the acceptability of the campaign arguments with respect to a standard argumentation semantics. In both cases, we identify some sufficient conditions for monotonicity and provide a preliminary discussion about their relevance and applicability in practice.

Keywords

Public interest communication, Computational argumentation, Value-based argumentation

1. Introduction

Public Interest Communication (PIC) involves the design and execution of evidence-based communication campaigns aimed at generating significant and sustained behavioural change on issues that serve the public good. Unlike communications driven by corporate or political interests, its primary goal is to advance causes that transcend the specific interests of any single organisation. It plays a vital role in encouraging beneficial behaviours and policies, clarifying their rationale, and securing legitimacy among key stakeholders, including institutions and the general public. To achieve these goals, Public Interest Communication relies primarily on carefully crafted communication campaigns that address specific target audiences. The overarching objective of a campaign is to persuade an audience to adopt a behavior by providing *arguments* for it, be it in the form of simple slogans or more articulated chains of reasoning to a conclusion.

Campaigns are designed with the assumption that effective messaging can lead to persuasion, ultimately resulting in behavioural change. However, the manner in which these messages resonate with individuals can vary significantly. The effectiveness of a campaign depends on the nature of its persuasive approach, which can take different forms. A rational approach appeals to logic and evidence, presenting factual information to support a particular stance. A value-based approach connects instead with the audience's deeply held beliefs and ethical considerations, reinforcing alignment between personal values and recommended behaviours. Emotional appeals, on the other hand, leverage feelings such as empathy, fear, or hope to create a strong psychological impact that motivates action. These approaches are not mutually exclusive and are often used in combination to maximise impact. For instance, a campaign promoting a greener diet might use multiple arguments to encourage behavioural change, each one leveraging on a different value or emotional trigger:

– Eating more fruits and vegetables contributes to better personal health.

AI[3] 2025: 9th Workshop on Advances in Argumentation in Artificial Intelligence, September 13, 2025, Rende, Italy

pietro.baroni@unibs.it (P. Baroni); giulio.fellin@unibs.it (G. Fellin); massimiliano.giacomin@unibs.it (M. Giacomin); carlo.proietti@ilc.cnr.it (C. Proietti)
0000-0001-5439-9561 (P. Baroni); 0000-0002-3179-7521 (G. Fellin); 0000-0003-4771-4265 (M. Giacomin); 0000-0001-9044-9642 (C. Proietti)

- Increasing fruit and vegetable consumption supports animal welfare.
- A plant-based diet has a lower environmental footprint.
- Choosing locally sourced produce benefits the local economy.

Despite their importance, public interest campaigns often face significant challenges. Ineffectiveness or backfire effects, caused by poorly targeted communication, are common issues, as evidenced by numerous unsuccessful and costly campaigns [1, 2]. One key problem is that these campaigns often target a general audience with diverse knowledge, needs, values, and attitudes. Finding a one-size-fits-all strategy is difficult, which is why campaigns typically leverage multiple motivations, as illustrated by the arguments in the greener diet example. This also makes it challenging to analyse the reasons for a campaign's success or failure. In fact this issue is the major topic of interest in the emerging field of *Public Interest Communication studies* [3, 4, 5].

Since PIC campaigns are essentially made of arguments, it is natural to leverage on tools from formal argumentation [6] to analyse them. In a previous work [7] we tackled this problem by means of a vector-based extension of of *value-based abstract argumentation*. The main use of such framework consists of assessing the *impact* on a specific *audience* of a single or multi-argument campaign. In a nutshell, the impact is assessed in terms of the set of weighted values, more precisely *vectors of weighted values*, promoted by the arguments and their relevance for a specific audience.[1] As a first proposal, we implemented an impact measure satisfying the desiderata for being a *seminorm* (Section 3). Although being natural in many contexts, this choice assumes that, other things being equal, the more strongly an argument refers to some value, the higher will be its impact, independently of the audience. It was also implicitly assumed that adding more arguments to a campaign cannot decrease the effectiveness of the campaign itself.

These assumptions imply somehow that, in terms of argument impact, *the more the better*, an effect we will refer to, in what follows, as *monotonicity*. While monotonicity may hold in some contexts, there are also situations where the addition of arguments and an extended coverage of values can be harmful and lead to a less effective communication. In fact, theories of basic human values as [8] call attention to the fact that some values often are in contrast with others, and therefore adding more stress on some of them may not increase the impact of an argument with regard to an audience that does not align with such values. Along similar lines, the so called *paradox of choice* [9] highlights the fact that providing more alternatives may create confusion and harm decision making, against the beneficial impact of adding more arguments, and to the effect that sometimes *less is more*. Further, psychological evidence also stresses that any argument that contrasts with one audience's background knowledge (e.g. other arguments they are aware of) may trigger a *backfire effect* [10] (see [11] for an application to argumentative scenarios).

Capturing this kind of situations requires a nonmonotonic behavior to be represented by the adopted model. In this work we address this requirement by investigating modifications of the model introduced in [7] which encompass non-monotonicity and discussing their properties and application. As our main contribution we prove two results providing sufficient conditions under which the effect of campaigns is guaranteed to be monotonic in the number of arguments (Proposition 4.4, contra the Paradox of Choice) and is immune to backfire effects (Proposition 5.8).

The paper is organised as follows. Section 2 recalls the necessary background notions, while Section 3 reviews the monotonic model of argument impact proposed in [7]. Section 4 discusses the issue of enconmpassing nonmonotonic argument impact measures, while Section 5 extends the analysis to the case where campaign effects are assessed through argumentation framework. Finally, Section 6 concludes the paper.

[1] Since different audiences may ascribe different importance to different values, it follows that in such contexts the acceptability of arguments becomes audience-dependent.

2. Background

In this section we provide the necessary background on Dung's theory of abstract argumentation (with focus on the notion of acceptability) hinting at how to expand it into value-based argumentation so that it fits the purposes of our conceptual model.

2.1. Argumentation frameworks and acceptability of arguments

Dung's theory of abstract argumentation treats arguments as abstract entities, whose internal structure and properties are abstracted away, and focuses only on conflicts between them. The key notion is that of an argumentation framework, defined as follows:

Definition 2.1. An argumentation framework (AF) is a pair $AF = \langle A, \mathcal{R} \rangle$, where A is a set of arguments and $\mathcal{R} \subseteq (A \times A)$ is a binary relation on A.

When $(a, b) \in \mathcal{R}$ (also denoted as $a\mathcal{R}b$) we say that a *attacks* b. For a set $X \subseteq A$ and an argument $a \in A$ we write $a\mathcal{R}X$ if $\exists b \in X : a\mathcal{R}b$ and $X\mathcal{R}a$ if $\exists b \in X : b\mathcal{R}a$, and we denote the arguments attacking X as $X^- \triangleq \{a \in A \mid a\mathcal{R}X\}$ and the arguments attacked by X as $X^+ \triangleq \{a \in A \mid X\mathcal{R}a\}$.

The relation of attack is the basis for the evaluation of the acceptability of arguments, given that the conflict among them prevent to accept them all together. Acceptability is determined by *argumentation semantics*. Formally, an argumentation semantics σ specifies the criteria for identifying, for a generic AF, a set of *extensions*, where each extension is a set of arguments considered to be acceptable together. Given a generic argumentation semantics σ, the set of extensions prescribed by σ for a given framework AF is denoted as $\mathcal{E}_\sigma(AF)$.

Typical (minimal) criteria for a set of arguments constituting an extension are *conflict-freeness*, the absence of conflict between its members, and *self-defense* the capacity of attacking every external attacker. Definition 2.2 recalls these notions and the definition of the grounded semantics, which is the only one we use in this paper. For more details, the reader is referred to [12].

Definition 2.2. Let $AF = \langle A, \mathcal{R} \rangle$ be an argumentation framework, $a \in A$ and $X \subseteq A$. X is *conflict-free*, denoted as $X \in \mathcal{E}_{\mathrm{CF}}(AF)$, iff $X \cap X^- = \emptyset$. a is *acceptable* with respect to X (or a is defended by X) iff $\{a\}^- \subseteq X^+$. The function $F_{AF} : 2^A \to 2^A$ which, given a set $X \subseteq A$, returns the set of the acceptable arguments with respect to X, is called the *characteristic function* of AF. X is *admissible* (denoted as $X \in \mathcal{E}_{\mathrm{AD}}(AF)$) iff $X \in \mathcal{E}_{\mathrm{CF}}(AF)$ and $X \subseteq F_{AF}(X)$. X is the *grounded* extension (denoted as $X = GR(AF)$ or $X \in \mathcal{E}_{\mathrm{GR}}(AF)$) iff X is the least fixed point of F_{AF}.

2.2. Value-based argumentation

Value-based argumentation frameworks have been introduced in [13]. They add two relevant dimensions for our modelling: (i) the introduction of a set of values $\mathcal{V}$ referring to arguments and (ii) a set U of different audiences, the target subjects of our modelling, where each audience u ranks values in different ways, specified by a ranking $\prec_u$. Crucially, each audience $u \in U$ is associated with an argumentation framework $AF_u = \langle A, \mathcal{R}_u \rangle$, where an argument $a \in A$ *defeats* an argument b for audience u, denoted as $(a, b) \in \mathcal{R}_u$ if and only if

$$a\mathcal{R}b \text{ and } v(a) \not\prec_u v(b)$$

In words, an argument a defeats b for audience u only when a attacks b in the ordinary sense and the audience does not rank the value of b higher than that of a. The acceptability of arguments according to audience u can then be derived by applying an argumentation semantics σ to AF_u.

The approach summarized above, where each argument relates exactly to a single value and values are ordered differently by different recipients, is arguably the most immediate one for representing how a value dimension may be attached to argumentative discourse. Yet, there are at least three aspects that are worth rediscussing in the context of our analysis of PIC campaigns. First, arguments may refer to more than a single value, as soon as they pertain to an articulated conceptual framework, while there can also be arguments which are not associated to any value. Second, as stressed by different theories of human values [14, 15], an argument may be anchored to values *with different degrees of intensity* and this matters for the recipient. Third, also relating to these empirical findings in psychology, some values may be mutually incompatible: a strong appeal to one value may negatively impact the persuasive force of an argument for certain audiences.

Regarding the first aspect, a generalizing approach by [16] already allows arguments to refer to multiple values (or none), where the ordering relation is lifted accordingly to sets of arguments. However, besides other shortcomings of this approach, the second and third aspect are hard to deal with a simple ordering of elements, but rather require a numerical treatment. Bringing these desiderata together, it is natural to use vectors with weighted coordinates to handle reference to multiple values and with different intensity. We instead replace the simple ordering among values, by a measure of their impact on an audience, as resulting from the distribution of weights to the vector's coordinates. The formal details of this approach are recapitulated in Section 3.

3. A monotonic model of argument impact

In this section we recall, with some minor formal adjustments, the model introduced in [7], with particular reference to the argument impact measure.

We assume that given a domain[2] of interest (e.g., the promotion of a greener diet) there is a reference universe of potential arguments, denoted as $\mathcal{A}$ and a set of relevant values $\mathcal{V}$. Both sets are assumed to be finite. The key point of our approach is that each argument $a \in \mathcal{A}$ is characterized by the set of values it promotes, where an argument can promote several values to different extents.

Definition 3.1 (Space of values)**.** We assume that there is a finite set $\mathcal{V}$ of values with cardinality $|\mathcal{V}| = n$. Each value v is identified by a number in $1 \dots n$. The *space of values* is $V = [0, 1]^n$, where each dimension is associated with the corresponding value, so that each point of the space corresponds to an assignment of weights in $[0, 1]$ to the values. Given a vector $vv \in V$ its j-th element will be denoted[3] as vv_j for $1 \leq j \leq n$. Further, we define a *value function* $\mathrm{val} : \mathcal{A} \to V$, which assigns each $a \in \mathcal{A}$ to its *vector of values*.

Arguments in the universe may attack each other, giving rise to a universal argumentation framework for the considered domain.

Definition 3.2 (Universal argumentation framework)**.** Given a universe $\mathcal{A}$ of potential arguments for a given domain, we assume the existence of a binary attack relation $\mathcal{R}^{po} \subseteq \mathcal{A} \times \mathcal{A}$. The universal argumentation framework, denoted as $\mathcal{AF}$, for the domain is defined as $\mathcal{AF} = \langle \mathcal{A}, \mathcal{R}^{po} \rangle$.

The attack-relation $\mathcal{R}^{po}$ is assumed to be subject-independent. When $(a, b) \in \mathcal{R}^{po}$ we say that a is a potential attacker of b.

In each domain, there is a set $\mathcal{S}$ of audiences, representing the potential targets of communication campaigns. Each audience $s \in \mathcal{S}$ will have their own *preferences* among values. Based on the discussion of Section 2, we represent this associating a profile of weights (one for each value coordinate) with each audience.

Definition 3.3. Given a set of audiences $\mathcal{S}$ and a space of values V, an *audience specific value function*

$$\mathrm{asv} : \mathcal{S} \to V$$

assigns to each audience $s \in \mathcal{S}$ a vector $\mathrm{asv}(s) \in V$ whose j-th entry represents the importance that the audience s assigns to value j.

[2] To avoid a too heavy notation, in the following we leave implicit the domain of interest $\mathcal{D}$, as it is unique for each application context. For instance we indicate the universe of arguments as $\mathcal{A}$ rather than as $\mathcal{A}_{\mathcal{D}}$ or any other notation evidencing the connection with $\mathcal{D}$.

[3] In the following we will use this notation to identify the j-th element of any kind of vector occurring in the paper.

Besides having preferences on values, each audience has an initial mindset on the considered domain, represented by the arguments in the universe which are known to the audience before a communication campaign is started.

Definition 3.4. Given a set of audiences $\mathcal{S}$ and a a universe $\mathcal{A}$ of potential arguments, the set of arguments initially known to each audience $s \in \mathcal{S}$ is denoted as $A_s^0 \subseteq \mathcal{A}$.

A campaign in a given domain consists of one or more arguments which are communicated to a selected set of audiences representing the target of the campaign. We assume that each campaign has a goal and that, on the basis of the goal, the subset of the arguments which are eligible for the campaign can be identified.

Definition 3.5. A PIC campaign is characterized by a goal g, whose nature is left unspecified. We assume that for each possible goal g a set $A_g \subseteq \mathcal{A}$ of relevant arguments is identified. The set of all possible PIC campaigns for a goal g, denoted as $\mathcal{C}^g$ consists of all non-empty subsets of A_g, namely $\mathcal{C}^g = \{S \subseteq A_g \mid S \neq \emptyset\}$. Given a campaing $c \in \mathcal{C}^g$ the corresponding set of arguments is denoted as $args(c)$. In the following, with a little abuse of notation, we will sometimes equate a campaign with its set of arguments where this is not ambiguous.

An *impact measure* is meant to give an account of how much each argument is effective for a given audience. In [7], this is expressed by a function $\|\cdot\|^s : \mathcal{A} \to [0,1]$ for each audience $s \in \mathcal{S}$. Intuitively the measure $\|\cdot\|^s$ should assign to each argument a an impact based on the values on which a relies and the importance that these values have for s. As a requirement, we need to ensure that if an argument a relies on one or more values that are important to s, then a will have a high impact on s and $\|a\|^s$ indicates the degree of acceptability of a according to s. In this spirit, $\|\cdot\|^s$ can be written as a composition

$$\|\cdot\|^s : \ \mathcal{A} \xrightarrow{\mathrm{val}} V \subseteq \mathbb{R}^n \xrightarrow{\mathrm{n}_s} \mathbb{R}$$

In words, given an argument a, its impact on audience s is determined on the basis of its value of vectors $\mathrm{val}(a)$ which is then synthesised into a single real number, measuring the impact, through an audience specific function n_s. In particular, in [7] the use of monotonic seminorms was advocated and, more specifically, the Euclidean norm was adopted, leading to $\|\cdot\|^s : \mathcal{A} \to [0,1]$ such that for every $a \in \mathcal{A}$:

$$a \mapsto \sqrt{\frac{1}{n}\sum_{j=1}^{n} \left(\mathrm{asv}(s)_j \cdot \mathrm{val}(a)_j\right)^2}. \tag{1}$$

In words, the impact here is calculated by multiplying each value coordinate of the given argument by the weight of this coordinate for the specific audience, and then by taking the average.

In order to model the impact of a campaign consisting of several arguments and directed to a set of audiences, some further notions have to be introduced. First we define the set of possible targets of a campaign, which need to be associated with a weighting, since different audiences may matter to different degrees with respect to the goal of the campaign (e.g. improving dietary habits can be more important for younger than for elder people).

Definition 3.6. A set TS of target audiences for a campaign is any non-empty subset of the set of audiences. Given a set TS representing a set of target audiences, we assume that each audience $s \in TS$ is assigned a weight $w(s)$, representing its importance in the campaign, which satisfies the following constraints.

$$\forall s \in TS, w(s) \geq 0, \qquad \sum_{s \in TS} w(s) = 1.$$

The overall impact of a campaign for a given set of target audiences can then be defined as a weighted sum of the impact, on each audience, of the arguments it contains[4].

Definition 3.7. Given a campaign c for a goal g consisting of a set of arguments $args(c) \subseteq A_g$ and a set TS of target audiences, the overall impact of c with respect to TS, denoted as $\mathcal{OI}(c, TS)$ is defined as

$$\mathcal{OI}(c, TS) = \sum_{s \in TS} w(s) \cdot \sum_{a \in args(c)} \|a\|^s \tag{2}$$

The model of impact summarized above is monotonic in two respects: with respect to the values promoted and with respect to the arguments included in the campaign.

As to the former aspect, given two value vectors $vv^1, vv^2 \in V$, assume the following relation: $vv^1 \preceq vv^2$ iff for every $1 \leq i \leq n$ $vv^1_i \leq vv^2_i$. Then, it is easy to see that assuming the use of equation (1), and indeed of any monotonic seminorm n_s, for any audience s it holds that if for two arguments a and b $\mathrm{val}(a) \preceq \mathrm{val}(b)$, then $\|a\|^s \leq \|b\|^s$. In words, the more an argument is able to promote all values the higher is its impact, or to say it compactly *promoting more values cannot do harm.*

As to the second aspect, the purely additive model encompassed by Definition 3.7 is based on the idea that each argument included in the campaign represents a further opportunity to convince the target audiences and hence, to say it compactly *making more articulated campaigns cannot do harm.*

Both aspects of monotonicity can be reasonable in some contexts, but fail to capture phenomena like backfire effects and would correspond to a rather simple criterion for campaign design: include as many arguments as possible promoting as many values as possible. The task of campaign design in practice is typically much more complex than this criterion would suggest. In next section, we discuss therefore alternative modeling options encompassing nonmonotonic impact measures.

4. Beyond monotonicity in argument impact

The measure introduced in equation (1) captures the intuition that the contribution of each value to the impact is non-negative. The idea is that if an audience s does not care about a value v (e.g. the relevant weight in $\mathrm{asv}(s)$ is zero), it will be essentially indifferent to the presence of v among the values promoted by an argument a, but this will not affect negatively the impact determined by the other values promoted by a. While this is reasonable in some contexts, there are also situations where one can imagine that the impact of an argument depends on *how close* the values promoted by the argument are to the values which are important for the audience. This means that the presence of a value which is not shared by the audience can affect negatively the impact of the relevant argument, possibly countering the role of other values. This may occur, for instance, in the cases where an audience has some strong biases and/or radical opinions and is inclined to reject every communication which, even partially, bears some similarity with positions perceived as "opposite" in some sense.

To represent this kind of situations, a nonmonotonic behavior of the argument impact measure is required, such that an audience can react negatively to the promotion of some values. We consider here the option of assessing impact in terms of the distance between the value vectors of the argument and of the audience. This leads to consider a combined vector whose elements result from the element-wise difference (in absolute value) rather than product of the value weights. This resorts to assessing the *distance* between the value vector the argument refers to and the audience specific value function. The impact of the combined vector can, in general, be measured through a norm, as seen in Section 3.

In general, a family of alternative distance measures can be considered. We assume that the impact is the complement to 1 of the distance between the value vector of the argument and the one of the audience. In words, if the two vectors coincide the impact is 1, otherwise it decreases with the distance

[4] This is a very simple instance of aggregation method, the consideration of alternative aggregation functions like min, max and, more generally, OWA operators [17] is left to future work.

and becomes 0 if the two vectors contain the opposite extremes (i.e. one is 0 and the other is 1 or vice versa) for each value. This can be captured by a distance-based function $\| \cdot \|^{ds} : \mathcal{A} \rightarrow [0, 1]$, which, assuming again the Euclidean case, can be defined as follows:

$$\| \cdot \|^{ds} = 1 - \sqrt{\frac{1}{n} \sum_{j=1}^{n} (\mathrm{val}(a)_j - \mathrm{asv}(i)_j)^2}. \tag{3}$$

Equation 3 overcomes the first kind of monotonicity discussed at the end of Section 3. Here, promoting more values, namely increasing some element of $\mathrm{val}(a)$ can be harmful since it may increase the distance term which is subtracted to 1 and hence may decrease the impact measure.

Turning to second kind of monotonicity, a further consideration concerns the possible negative aspects of multiplicity. In the model introduced in Section 3, potentially every addition of a further argument to a campaign can increase its collective impact. However, this clashes against the intuition that long and verbose messages are often less effective than concise and focused ones. Moreover, if there are multiple items in a message, it is common that they do not receive the same attention, as (some) people more easily focus on the first (and possibly the last) parts of a communication while it is more likely that they overlook the contents in the middle.

To provide a formal counterpart to this intuition we need a small modelling refinement, namely that the arguments forming a campaign c for a goal g, i.e. the arguments in $args(c)$, are arranged in an ordered list denoted as $\mathcal{L}(c) \in A_g^*$, where, for a set S, S^* denotes the set of all (finite) sequences of elements of S. For $a \in args(c)$ the position of a in the list $\mathcal{L}(c)$ will be denoted as $pos^c(a)$. Then, we assume that each audience has its own capabilities and attitudes concerning the reception of ordered communications of a given length and that, in this respect, the effectiveness of each item is affected by its position in the list. To represent this aspect we associate to each audience s a position-weighting function $O_s : \mathbb{N} \rightarrow [0, 1]^*$ which for each natural number n (i.e. for each possible list length) returns a vector of length n, consiting of the positional weights of audience s for a list of length n. In words, the j-th element of the vector is the weight with which audience s receives the j-th item in an ordered communication.

The considerations above lead to revise Definition 3.7 as follows.

Definition 4.1. Given a campaign c for a goal g, with argument list $\mathcal{L}(c)$ of cardinality n, and a set TS of target audiences, the position-aware overall impact of c with respect to TS, denoted as $\mathcal{PI}(S, TS)$ is defined as

$$\mathcal{PI}(c, TS) = \sum_{s \in TS} w(s) \cdot \sum_{a \in args(c)} \|a\|^s \cdot O_s(n)_{pos^c(a)} \tag{4}$$

Definition 4.1 generalises Definition 3.7, which can be regarded as a special case where for every audience s $O_s(n)$ returns a vector consisting of n 1's. We can make some comments on the nature $O_s(n)$ and how it gives rise to nonmonotonicity of the impact measure.

First, it seems rather reasonable, though the model allows also other choices, that for every audience $O_s(1) = \langle 1 \rangle$, i.e. that being the only argument used in a campaign cannot be detrimental to the impact of the argument *per se*, which in this case is at the same time in the first position, in the last one and (in a degenerate sense) in all the intermediate positions. Peculiar exceptions to this assumptions can however be conceived. For instance, if one assumes an audience whose attention needs to be "warmed up", a single argument may not be received with full weight, this being reserved to arguments occurring after a given position.

Second, it seems reasonable that the greater is n the higher is the chance that $O_s(n)$ contains some zeros or anyway some very low values. This can be intuitively related to an excess of cognitive load, which can manifest itself either in forgetting the first arguments or losing attention on the last ones (or both) leading respectively to having zeros at the beginning or at the end of the vector $O_s(n)$ for a large n. It is also interesting to note that different audiences may have different attention capabilities or cognitive load thresholds. These can be put in correspondence with the value of the sum of the elements of $O_s(n)$ with the increase of n. For instance, if for a given n and two audiences s_1 and s_2 it holds that $\sum_{i=1}^{n} O_{s_1}(n)_i < \sum_{i=1}^{n} O_{s_2}(n)_i$, this means that the audience s_2 is altogether more capable to sustain the cognitive load of receiving n arguments.

While this model is admittedly rough (for instance the cognitive load may depend not only on the number of the arguments but also on their structure or form of presentation) it allows to capture a variety of possible situations and in particular to introduce nonmonotonicity with respect to the set of arguments included in the campaign, even if each argument would *in isolation* have a nonnegative impact on all audiences. In particular, it may be the case that for some audience s it holds that $\sum_{i=1}^{n_1} O_s(n_1)_i < \sum_{i=1}^{n_2} O_s(n_2)_i$ for some n_1, n_2 such that $n_1 > n_2$. This would correspond to a situation where a too long list affects negatively the impact of all the arguments it contains and a shorter list would be more effective.

In general, it can be interesting to identify conditions on O_s ensuring that the effectiveness of a campaign cannot decrease by adding new arguments. To draw considerations in this respect we need to impose some constraints on the order of the arguments in the initial and in the extended campaign. In particular we assume that the arguments with greater impact (i.e. arguments with higher values of $\|a\|^s$) are put in the positions with greater positional weight. This is expressed by the notion of *position-optimal* campaign.

Definition 4.2. A campaign c for a goal g, with argument list $\mathcal{L}(c)$ of cardinality n is position-optimal for an audience s if for every pair of arguments $a, b \in \mathit{args}(c)$ it holds that $\|a\|^s \leq \|b\|^s$ if and only if $O_s(n)_{pos^c(a)} \leq O_s(n)_{pos^c(b)}$. A campaign c is position-optimal for a set of audiences if and only if it is position-optimal for all the audiences in the set.

We can now establish a simple non-decreasing condition for O_s.

Definition 4.3. Given $n_1, n_2 \in \mathbb{N}$ with $n_1 < n_2$, we say that a position-weighting function O_s of an audience s is non-decreasing with respect to n_1 and n_2 iff there is an injective function f from the set $\{1, \ldots, n_1\}$ to the set $\{1, \ldots, n_2\}$ such that for each $1 \leq i \leq n_1$ it holds that $O_s(n_1)_i \leq O_s(n_2)_{f(i)}$.

In words, for every position from 1 to n_1 there is a distinct position included between 1 and n_2 which has a not lesser positional weight.

A monotonicity guarantee follows from the above conditions.

Proposition 4.4. *Let c_1, c_2 be two campaigns such that $\mathit{args}(c_1) \subsetneq \mathit{args}(c_2)$ and let TS be a set of target audiences. If for every audience $s \in TS$ it holds that c_1 and c_2 are position optimal and O_s is non-decreasing with respect to $|\mathit{args}(c_1)|$ and $|\mathit{args}(c_2)|$ it follows that $\mathcal{PI}(c_1, TS) \leq \mathcal{PI}(c_2, TS)$.*

Proof. First note that all terms in the sum defined in (4) are nonnegative. Then observe that for every audience in TS, every argument in $\mathit{args}(c_1)$ contributes to $\mathcal{PI}(c_1, TS)$ with a term $\|a\|^s \cdot O_s(n)_{pos^{c_1}(a)}$ (and similarly for every argument in $\mathit{args}(c_2)$). Let $a_1, \ldots, a_{n_1}$ be any ordering of the arguments in $\mathit{args}(c_1)$ such that $\|a_1\|^s \geq \|a_2\|^s \ldots \geq \|a_{n_1}\|^s$. Similarly let $b_1, \ldots, b_{n_2}$ be any ordering of the arguments in $\mathit{args}(c_2)$ such that $\|b_1\|^s \geq \|b_2\|^s \ldots \geq \|b_{n_2}\|^s$. To show that $\mathcal{PI}(c_1, TS) \leq \mathcal{PI}(c_2, TS)$ we show that for every $1 \leq i \leq n_1$ $\|a_i\|^s \cdot O_s(n)_{pos^{c_1}(a_i)} \leq \|b_i\|^s \cdot O_s(n)_{pos^{c_2}(b_i)}$. Consider first a_1: since there is some j such that $a_1 = b_j$, it follows that $\|b_1\|^s \geq \|a_1\|^s$, moreover from the fact that both campaigns are position optimal we get that $O_s(n)_{pos^{c_1}(a_1)} \geq O_s(n)_{pos^{c_1}(a_k)}$ for every $1 \leq k \leq n_1$ and similarly $O_s(n)_{pos^{c_2}(b_1)} \geq O_s(n)_{pos^{c_2}(a_k)}$ for every $1 \leq k \leq n_2$. From the hypothesis that O_s is non-decreasing with respect to $|\mathit{args}(c_1)|$ and $|\mathit{args}(c_2)|$ it follows that $O_s(n)_{pos^{c_1}(a_1)} \leq O_s(n)_{pos^{c_2}(b_1)}$ from which $\|a_1\|^s \cdot O_s(n)_{pos^{c_1}(a_1)} \leq \|b_1\|^s \cdot O_s(n)_{pos^{c_2}(b_1)}$. Moving to a_2, from the fact that $a_1 = b_j$ for some $j \geq 1$ and that $a_2 = b_k$ for some $k > j$ it follows that $\|b_2\|^s \geq \|a_2\|^s$. From the fact that both campaigns are position optimal we get that $O_s(n)_{pos^{c_1}(a_2)} \geq O_s(n)_{pos^{c_1}(a_k)}$ for every $2 \leq k \leq n_1$ and similarly $O_s(n)_{pos^{c_2}(b_2)} \geq O_s(n)_{pos^{c_2}(a_k)}$ for every $2 \leq k \leq n_2$. From the hypothesis that O_s is non-decreasing with respect to $|args(c_1)|$ and $|args(c_2)|$ it follows that $O_s(n)_{pos^{c_1}(a_2)} \leq O_s(n)_{pos^{c_2}(b_2)}$ from which $\|a_2\|^s \cdot O_s(n)_{pos^{c_1}(a_2)} \leq \|b_2\|^s \cdot O_s(n)_{pos^{c_2}(b_2)}$. The same reasoning can be iterated for all arguments until a_{n_1} reaching the desired conclusion. □

The study of other sufficient conditions for monotonicity and of how the above identified condition can be reasonably met in practice is left to future work. As preliminary comments we can observe that if TS consists of a single audience s, assuming that $\| \cdot \|^s$ and O_s are known, it is rather reasonable to assume that one designs a campaign which is position optimal for s. If TS consists of a different audiences it may be difficult, and sometimes provably impossible, to define a campaign which is position optimal for all audiences. As to the requirement of non-decreasing position-weighting functions, it appears to be heavily audience dependent, and, in general, to correspond to a sort of robustness of the attention and reception capabilities of the audience with respect to the increase of the cognitive load. Leaving the investigation of these aspects to further research, in the sequel of the paper we address the issue of monotonicity of campaign effects from the perspective of argument acceptability.

5. Assessing campaign effects through argumentation frameworks

In [7], in addition to introducing a vector-based approch for the quantitative assessment of the impact of a communication campaign, an alternative perspective, based on the use of argumentation frameworks has been considered. Here we recall and integrate this perspective and discuss its relationships with monotonicity.

The starting assumption is that it is possible to identify, for each audience s, an argumentation framework, denoted as AF_s^0, representing the mental state of the audience before receiving the campaign. In particular, the arguments in AF_s^0 are a subset of the universe of arguments initially known to the audience (Definition 3.4), and the attacks are derived from those in the universal argumentation framework (Definition 3.2) following the idea of the value-based approach as specified in the following definition.

Definition 5.1. Given an audience s, the initial argumentation framework of s is defined as $AF_s^0 = \langle A_s^0, \mathcal{R}_s^0 \rangle$ where $A_s^0 \subseteq \mathcal{A}$ and $\mathcal{R}_s^0 = \{(a, b) \in A_s^0 \times A_s^0 \mid (a, b) \in \mathcal{R}^{po} \text{ and } \|a\|^s \not< \|b\|^s\}$.

After receiving a campaign c each audience updates its argumentation framework. In particular, we assume the existence of an audience-specific attention-trigger map which specifies, given a set of arguments S, which arguments are brought to the attention of the audience after being exposed to the arguments in S. As a minimal requirement, we assume that at least the arguments included in the campaign are brought to the attention of the audience.

Definition 5.2. Given an audience s, an attention-trigger map for s, denoted as $\mathcal{AT}_s$, is a function $\mathcal{AT}_s : 2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$, satisfying the requirement that for every $S \subseteq \mathcal{A}$ it holds that $S \subseteq \mathcal{AT}_s(S)$.

The argumentation framework of the audience after receiving a campaign includes also the arguments triggered by the campaign and the relevant attacks, again following the value-based perspective.

Definition 5.3. Given an audience s and a campaign c, the argumentation framework of s after receiving c is defined as $AF_s^c = \langle A_s^c, \mathcal{R}_s^c \rangle$, where $A_s^c = A_s^0 \cup \mathcal{AT}_s(args(c))$ and $\mathcal{R}_s^c = \{(a, b) \in A_s^c \times A_s^c \mid (a, b) \in \mathcal{R}^{po} \text{ and } \|a\|^s \not< \|b\|^s\}$

For simplicity, we assume here the use of a single-status semantics, namely the grounded semantics[5], for the assessment of acceptance of the arguments in an argumentation framework. An audience s meets the goal g of a campaign if at least one of the arguments in A_g is accepted by s in the updated argumentation framework. To formalize this we resort to the following definition.

[5] The consideration of other semantics is left to future work.

Definition 5.4. Given a campaign c with goal g and an argumentation framework AF, we say that the goal of the campaign is met with respect to AF, denoted as $OK(c, AF)$, iff $A_g \cap GR(AF) \neq \emptyset$.

Note that of course it can happen that an audience meets the goal of the campaign even before receiving it, namely it can happen that $OK(c, AF_s^0)$. It follows that with respect to the effects of a campaign we can identify three cases.

Definition 5.5. Given a campaign c and an audience s we say that

- s is positively affected by c, denoted as $E^+(s, c)$ iff $OK(c, AF_s^c)$ while not $OK(c, AF_s^0)$;
- s is negatively affected by c, denoted as $E^-(s, c)$ iff $OK(c, AF_s^0)$ while not $OK(c, AF_s^c)$;
- s is unaffected by c otherwise.

The assessment of a campaign can then consist of the sum of the weights of the audiences on which it has a positive effect from which the weights of the audiences on which it has a negative effect should be subtracted.

Definition 5.6. Let c be a campaign and TS be a set of target audiences. The conviction value of c with respect to TS is defined as

$$CV(c, TS) = \sum_{s \in TS | E^+(s,c)} w(s) - \sum_{s \in TS | E^-(s,c)} w(s) \tag{5}$$

Some remarks on Definition 5.6 are worth making. First, differently from the assessment based on impact measures, it allows in principle also to derive a negative evaluation for a campaign, corresponding to situations where it has prevailing unintended effects. This is a significant improvement of the expressiveness of the model, which comes at the price of requiring the (non trivial in practice) capability of estimating not only the attitudes of the audiences with respect to values but also the arguments they initially hold and those which are triggered by the campaign. Second, it adopts a binary notion of acceptance and hence of success (and unsuccess) thus allowing a simpler analysis and discussion at this preliminary level of investigation. A finer representation could encompass some gradual notion, e.g. counting the number of accepted arguments in A_g as a non-binary measure of success or distinguishing the status of the arguments which are attacked by the grounded extension from the status of those which are not included in the grounded extension but are not attacked by it, i.e. are in a sort of undecided, rather than definitely rejected, situation. Leaving these developments to future work, let us now discuss monotonicity issues concerning the conviction values of campaigns.

Concerning the first kind of monotonicity discussed in Section 4, namely the one referring to the role of values, we observe that in the context of the assessment introduced in this section, values play an indirect role. In fact, while $\|a\|^s$ was a term directly affecting the outcome of the assessment both in Definition 3.7 and Definition 4.1, here its role consists in determining which attacks are effective. In particular, if one assumes that a campaign only consists of sensible arguments which should be accepted by the audiences, it is generally preferable that for every argument a in the campaign $\|a\|^s$ is as high as possible. Then, concerning the way $\|a\|^s$ is calculated, no additional considerations with respect to those drawn in previous sections are needed.

Concerning, the second aspect of monotonicity, namely the inclusion of more arguments in a campaign, the discussion is more articulated and partly tricky. A first key point concerns the attention-trigger map. Besides the basic assumption that a set of argument triggers at least all its elements, there are no other hypotheses on the function $\mathcal{AT}_s$, which in principle might "activate" for some audience s also arguments which are fully in contrast with the goal of the campaign and which might be in line with the values of the audience, thus possibly resulting in getting s to be negatively affected by c and then in a decrease of $CV(c, TS)$. It follows that any consideration on monotonicity must refer to some additional hypothesis on $\mathcal{AT}_s$. In particular, we resort to the assumption that the campaign is goal-coherent, namely that it does not trigger any argument which is in possibly indirect conflict with any argument in A_g.

Definition 5.7. Given a campaign c with goal g we say that c is goal-coherent with respect to an audience s if for every $a \in \mathcal{AT}_s(args(c))$ and every $b \in A_g$ it is not the case that a is an indirect attacker of b.

If a campaign is goal-coherent it cannot have negative effects.

Proposition 5.8. *Let s be an audience and c a goal-coherent campaign with goal g, then s cannot be negatively affected by c.*

Proof. To prove the statement we show that if there is some argument $a \in A_g$ such that $a \in GR(AF_s^0)$ then it also holds that $a \in GR(AF_s^c)$.

Given the well-known properties of grounded semantics (see [18] and the notion of strong defense in [19]), an argument a in an argumentation framework $AF = \langle A, \mathcal{R} \rangle$ belongs to the grounded extension $GR(AF)$ iff a is strongly defended in AF namely (i) a is unattacked in AF or[6] (ii) for every attacker b of a there is an attacker c of b such that $c \neq a$ and c is in turn strongly defended in AF.

By hypothesis, the above conditions hold for some argument $a \in A_g$ in the argumentation framework AF_s^0. We prove then that they hold also in AF_s^c. Note that any additional argument in AF_s^c with respect to AF_s^0 is included in $\mathcal{AT}_s(args(c))$ and that any additional attack must necessarily involve an element of $\mathcal{AT}_s(args(c))$. Now, if $a \in A_g$ is unattacked in AF_s^0, it is also unattacked in AF_s^c: any new attacker should belong to $\mathcal{AT}_s(args(c))$, which would contradict the property of goal coherence.

If instead a has a non-empty set of attackers in AF_s^0, we observe that, for the same reason, the set of attackers of a in AF_s^c is the same. Considering any attacker b of a: we have to show that the condition (ii) still holds in AF_s^c. Given that $\mathcal{R}_s^0 \subseteq \mathcal{R}_s^c$ it still holds in AF_s^c that $c \neq a$ attacks b. Suppose by contradiction that c is no more strongly defended in AF_s^c. For this to happen, c must have some additional indirect attacker d in AF_s^c, but then d would be also an indirect attacker of a and d should be a member of $\mathcal{AT}_s(args(c))$, which would contradict the property of goal coherence.

□

We have thus provided a sufficient condition to ensure that adding arguments does not affect negatively a campaign. The condition reflects a quite reasonable requirement of absence of conflicts (even indirects) with the arguments related to the campaign goal. The tricky point is that this requirement does not refer only to the arguments included explicitly in the campaign but also to those possibly triggered by the campaign itself for all the audiences included in the target set. In turn, this means that it is preferable that the audiences in the target set are somehow similar as far as the triggered arguments are concerned. In general, the problem is to achieve a sufficiently accurate estimation of the triggered arguments: in this respect, investigating triggering mechanisms (e.g. by analogy, by opposition, by inference, ...) appears a very relevant direction of future work.

6. Conclusions

Defining formal models able to support the design and evaluation of public interest communication campaigns is a challenging research goal, whose complexity suggests to follow an incremental approach, by addressing step by step its many facets. Building on a previous work investigating the definition of impact measures of argumentative communication campaigns based on the values they promote, in this paper we have made a further step by considering nonmonotonic campaign effects, i.e. situations where increasing the set of promoted values and/or extending a campaign with additional arguments can turn out to be detrimental, rather than beneficial, to the campaign itself.

We addressed nonmonotonicity both in quantitative impact measures directly assessing the effectiveness of a campaign and in evaluations of audience convictions based on argumentation frameworks. In both cases, we provided sufficient conditions to ensure monotonicity. This preliminary study provides the basis for further research on this topic. In addition to the future work directions indicated along the

[6] Condition (i) is actually a special case of condition (ii) but we keep it distinct for the sake of readability.

paper we mention the consideration of bipolar argumentation frameworks [20] to represent audience opinions and the use of gradual argumentation semantics [21], encompassing a quantitative notion of degree of acceptance.

Moreover, it would be interesting to investigate relationships between the proposed model and cognitive theories of persuasion [22] in order to justify the choice of alternative impact measures with some theoretical background or experimental evidence concerning the attitudes of different social categories. The model will then need to be validated using data concerning past PIC campaigns, in order to check the ability of our proposal to provide a model-based explanation of their successes or failures.

Acknowledgments

This work was supported by MUR project PRIN 2022 EPICA "Enhancing Public Interest Communication with Argumentation" (CUP D53D23008860006) funded by the European Union - Next Generation EU, mission 4, component 2, investment 1.1.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] R. Rekhy, R. McConchie, Promoting consumption of fruit and vegetables for better health. have campaigns delivered on the goals?, Appetite 79 (2014) 113–123.

[2] R. Hornik, L. Jacobsohn, R. Orwin, A. Piesse, G. Kalton, Effects of the national youth anti-drug media campaign on youths, American Journal of Public Health 98 (2008) 2229–2236.

[3] J. Fessmann, The emerging field of public interest communications, in: Strategic communications for non-profit organizations, Vernon Press, 2016, pp. 13–33.

[4] A. Christiano, Building the field of public interest communications, J. of Public Interest Communications 1 (2017) 4–15.

[5] K. Demetrious, Public Interest Communication, Oxford Bibliographies, 2023.

[6] P. Baroni, D. Gabbay, M. Giacomin, L. van der Torre (Eds.), Handbook of Formal Argumentation, College Publications, 2018.

[7] P. Baroni, G. Fellin, M. Giacomin, C. Proietti, A vector-based extension of value-based argumentation for public interest communication, in: C. Proietti, C. Taticchi (Eds.), Proc, of the 8th Workshop on Advances in Argumentation in Artificial Intelligence, volume 3871 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024.

[8] S. H. Schwartz, Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries, Advances in Experimental Social Psychology 25 (1992) 1–65.

[9] B. Schwartz, The paradox of choice, Positive psychology in practice: Promoting human flourishing in work, health, education, and everyday life (2015) 121–138.

[10] B. Nyhan, J. Reifler, When corrections fail: The persistence of political misperceptions, Political Behavior 32 (2010) 303–330.

[11] C. Proietti, D. Chiarella, The role of argument strength and informational biases in polarization and bi-polarization effects, Journal of Artificial Societies and Social Simulation 26 (2023).

[12] P. Baroni, M. Caminada, M. Giacomin, An introduction to argumentation semantics, Knowledge Engineering Review 26 (2011) 365–410.

[13] T. J. M. Bench-Capon, Persuasion in practical argument using value-based argumentation frameworks, J. Log. Comput. 13 (2003) 429–448.

[14] S. H. Schwartz, J. Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J.-E. Lönnqvist, K. Demirutku, O. Dirilen-Gumus, M. Konty, Refining the theory of basic individual values, Journal of Personality and Social Psychology 103 (2012) 663–688.

[15] J. Haidt, C. Joseph, Intuitive ethics: How innately prepared intuitions generate culturally variable virtues, Daedalus 133 (2004) 55–66.

[16] S. Kaci, L. W. N. van der Torre, Preference-based argumentation: Arguments supporting multiple values, Int. J. Approx. Reason. 48 (2008) 730–751.

[17] R. R. Yager, J. Kacprzyk, The Ordered Weighted Averaging Operators: Theory and Applications, Kluwer, 1997.

[18] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artif. Intell. 77 (1995) 321–358.

[19] P. Baroni, M. Giacomin, On principle-based evaluation of extension-based argumentation semantics, Artif. Intell. 171 (2007) 675–700.

[20] C. Cayrol, M. Lagasquie-Schiex, Bipolar abstract argumentation systems, in: G. R. Simari, I. Rahwan (Eds.), Argumentation in Artificial Intelligence, Springer, 2009, pp. 65–84.

[21] P. Baroni, A. Rago, F. Toni, From fine-grained properties to broad principles for gradual argumentation: A principled spectrum, Int. J. Approx. Reason. 105 (2019) 252–286.

[22] A. H. Eagly, S. Chaiken, Cognitive theories of persuasion, volume 17 of *Advances in Experimental Social Psychology*, Academic Press, 1984, pp. 267–359.