A formal framework for deliberated judgment*

Olivier Cailloux and Yves Meinard

Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, 75016 PARIS, FRANCE
olivier.cailloux@dauphine.fr

19th May, 2019

While the philosophical literature has extensively studied how decisions relate to arguments, reasons and justifications, decision theory almost entirely ignores the latter notions. In this article, we elaborate a formal framework in order to introduce in decision theory the stance that decision-makers take towards arguments and counter-arguments.We start from a decision situation, where an individual requests decision support. We formally define, as a commendable basis for decision-aid, this individual's deliberated judgment, a notion inspired by Rawls' contributions to the philosophical literature, and embodying the requirement that the decision-maker should carefully examine arguments and counter-arguments. We explain how models of deliberated judgment can be validated empirically. We then identify conditions upon which the existence of a valid model can be taken for granted, and analyze how these conditions can be relaxed. We then explore the significance of our framework for the practice of decision analysis. Our framework opens avenues for future research involving both philosophy and decision theory, as well as empirical implementations.

*This is the postprint version of an article to appear in Theory and Decision.

1. Introduction

Introducing their "reason-based theory of choice", Dietrich and List (2013) noticed that, although the philosophical literature has largely illustrated the usefulness of the concepts of reasons and arguments to think through action and decisions, decision theory strives to account for the latter exclusively in terms of preferences and beliefs. Despite Dietrich and List's (2013; 2016) efforts, the gap remains large between philosophical and choice theoretic approaches.

This gap echoes a classical dichotomy in "moral sciences" between, on the one hand, first-person justifications of one's acts in terms of reasons and arguments structuring these reasons, and on the other hand, third-person representations in terms of beliefs and preferences (Hausman, 2011). By neglecting reason-based and other argumentative accounts, decision theory tends to devalue decision-makers' understanding of their own actions.

This gap has tended to insulate decision theory from important philosophical debates in the past thirty to forty years. Among the most influential approaches in these debates, Scanlon (2000) highlighted the links between reasons, justification and moral notions such as fairness and responsibility, Habermas' (1981) "theory of communicative action" articulated the importance of justification and argumentation as distinctive features of rational action, and Rawls (2005) launched the debates on the "acceptability" (Estlund, 2009) of reasons and arguments for public justification.

This gap also has important practical implications for decision analysis, by complicating the task for analysts to explain the recommendations they give to their clients. This, in turn, casts doubts on these recommendations, which appear to be imposed to rather than endorsed by decision-makers.

In this article, we aim to participate in unlocking this situation, by elaborating a framework designed to allow decision analysts to provide recommendations that decision-makers truly endorse, in empirical reality.

For that purpose, we introduce, as a commendable basis for recommendation, the "deliberated judgments" of the decision-maker. Roughly stated, these "deliberated judgments" represent the propositions that the decision-maker will consider to be well-grounded, if he duly takes into account all the relevant arguments. This concept is inspired by Goodman's (1983) and Rawls' (1999) notion of reflective equilibrium. It also owes much to Roy's (1996) view that an important part of the decision support interaction consists, for the analyst, in ensuring that the aided individual understands and accepts the reasoning on which the prescription is based.

This article is organized as follows. In section 2, we define our core concepts, including the central concept of deliberated judgments. In section 3, we then

explore the issue of how empirical data come into play and are involved in the validation of models. This illustrates the empirical aspect of our framework, which distinguishes it from standard prescriptive approaches. Obviously enough, at this stage, the pivotal issue is to determine how one can say anything about "deliberated judgments", given that, for any non-trivial decision, the potentially relevant arguments are infinitely numerous. Lastly, section 4 discusses the significance of our approach for the practice of decision analysis and outlines future empirical applications.

2. Core concepts and notations

In this section, we start by presenting the general setting of our approach, including our understanding of arguments and of the topic on which the individual aims to take a stance. We then introduce our formalization of argumentative disposition, capturing an individual's attitude towards arguments. This eventually allows us to present our notion of "deliberated judgment".

2.1. General setting

Our approach starts from and is largely structured by the point of view of decision-analysis. We accordingly assume that a decision situation has been identified: we admit that there is an individual i who requests decision support to answer questions such as: "is action a better than action b?", or "which beliefs should I have about such or such matter?". We consider that a topic T – a set of propositions on which the decision analysis process aims to lead the decision-maker to take a stance – is defined.[1] We do not formally define propositions and simply understand the notion in its ordinary sense. For example, a proposition can be a claim spelled out in a text in a natural language, such as the claim that action a is the most appropriate action for i in a given decision situation.

We also consider arguments that can be used by i to make up her mind about propositions in T. Here we understand the notion of argument in a large sense: anything that can be used to support a proposition, or undermine the effectiveness of such a support, is an argument. In the latter case, we talk about a counter-argument. Arguments as we understand them can encompass

[1] We remain at a fairly abstract level in our conceptualization of the topic. We accordingly set aside all the issues concerning the construction of problems and the evolution of their meaning as the decision process unfolds in concrete decision situations (Rosenhead and Mingers, 2001).

a huge diversity, ranging from very basic arguments that can be stated in a couple of words, to intricate arguments embedding numerous sub-arguments associated to one another in complex ways.

Let us then define the set S^* that contains all the arguments that one uses when trying to make up one's mind about T. S^* can be understood in a "pragmatic" sense, as the set of all the arguments available around the temporal window of the decision process. It can also be understood in an "idealistic" sense, as the set of all the arguments that can possibly be raised, including those that humankind has not yet discovered.[2]

Observe that under both interpretations, in all decision situations but the most trivial, it will be untenable to assume that the analyst knows all of S^*: the analyst will only know a strict subset $S \subset S^*$, containing the arguments that she has been able to gather.[3] An important part of our work in this article will be to identify conditions allowing to draw conclusions relating to S^* despite the fact that no one ever knows more than a strict subset of S^*.

Example 1 (Ranking). Let us simply illustrate the content of the concepts introduced so far. Let $\mathcal{A}$ be a set of alternatives that i is interested in ranking. For all $a_1 \neq a_2 \in \mathcal{A}$, define $t_{a_1 \succ a_2}$ as the sentence: "a_1 ought to be ranked above a_2", and $t_{a_1 \sim a_2}$ as "a_1 ought to be ranked *ex-æquo* with a_2". Define $T = \bigcup_{a_1 \neq a_2 \in \mathcal{A}} \{t_{a_1 > a_2}, t_{a_1 \sim a_2}\}$ as the set of all such sentences. The topic T represents the propositions on which i is interested to make up her mind. Define S^* as the set of all strings corresponding to sentences in English. This set contains formulations of all the arguments that people can think about and use to make up their mind about the topic, and much more. An example of an argument is $s =$ "Alternative a_1 ought to be ranked above a_2 because a_1 is better than a_2 on every criterion relevant to this problem". △

Our aim in the remainder of this section is to define formally i's perspective towards the topic after he has considered all the arguments that are possibly relevant to the situation. We term this: i's Deliberated Judgment (DJ).

[2]Because no one has a concrete access to such an idealistic set of all the arguments, we expect that this concept will be mainly useful for philosophical explorations, and that the pragmatic interpretation will prevail in practical applications.

[3]Even under the pragmatic interpretation, claiming that $S = S^*$ would mean that there is no relevant knowledge beyond what the analyst can find by studying the literature and consulting experts and stakeholders, but also that the list of arguments she has found captures all the semantic and linguistic subtleties that could distinguish alternative formulations of arguments.

2.2. Argumentative disposition

To define i's DJ, we need to capture i's attitude towards arguments. Importantly, we also need to capture the fact that i may change her opinion about arguments and their relative strengths. She can change her mind because of reasons independent of her endeavor to tackle the problem she addresses, for example depending on her mood. More interestingly, i will possibly change her mind when confronted with new arguments. For example, imagine that i has heard about two arguments, s_1 and s_2, and she thinks that s_2 turns s_1 into an ineffective argument. But then she comes to realize that s_2 is in turn rendered ineffective by a third argument, s_3. After having thought about s_3, it might be that i no longer considers that s_2 undermines s_1.

Note that for simplicity's sake, we say that an argument becomes ineffective (because of another argument) to mean that it becomes ineffective in its ability to support some proposition or to render other arguments ineffective.

Le us introduce our formalism to account for such a situation.[4]

[4]Our approach to formalize this concept is inspired by formal argumentation theory in artificial intelligence (Dung, 1995; Rahwan and Simari, 2009). However, the latter approach is not sufficient to empirically investigate i's attitude towards arguments, because it neglects two crucial tasks. First, this literature does not investigate the role that the decision analyst plays when she interacts with a decision-maker: should she remain a neutral observer, or should she interact more tightly with the decision-maker by providing him with arguments and counter-arguments liable to lead him to change his mind? Second, this literature does not put emphasis on the specific challenges involved in interacting with a decision-maker to identify empirically the arguments he endorses. Most of the time, this literature considers situations where the relation between arguments can be computed from a given logical representation of the arguments (Besnard and Hunter, 2008) or is given *a priori* (Baroni and Giacomin, 2009), possibly integrating uncertainties (Hunter, 2014) and dynamics (Rotstein et al., 2010; Marcos et al., 2011; Dimopoulos et al., 2018).

Its most common use assumes that it is possible to establish the objective relations between arguments. In our example, s_3 would be considered to objectively attack s_2 and s_2 to objectively attack s_1. However, in some cases, it might be difficult, or perhaps even impossible, to determine such objective relations. In any case, this distinction is superfluous if the goal is to inquire about i's opinion about these relations between arguments. Other proposals in formal argumentation theory (Amgoud and Cayrol, 2002; Bench-Capon, 2003; Amgoud et al., 2008; Amgoud and Prade, 2009; Bench-Capon and Atkinson, 2009; Ferretti et al., 2017) supplement an objective attack relation with information representing i's subjectivity, such as his values or his preference over arguments. Such approaches seem closer to our aim, but they also use an objective attack relation, in addition to the subjective information. Furthermore, this approach assumes that it is possible to distinguish between, on the one hand, cases where s_3 attacks s_2 but i does not deem this attack important, and on the other hand situations where s_3 does not attack s_2. This assumption is also unnecessary for our purpose. Because our aim is mainly em-

Let us start by defining a set of possible perspectives P that i can have towards the topic T. A perspective $p \in P$ captures all the elements determining how i would react to arguments in S^*. In p, i has a specific set of arguments in mind, which can partly determine his reaction to other arguments in S^*. But other elements can come into play, such as (to come back to our example above) his mood.

If the decision analyst provides i with a new argument s, this might lead i to switch from p to another perspective p' integrating both s and the arguments that i had in mind in p, and possibly other arguments that i might have been led to construct when trying to make up his mind about s and its implications. i's perspective can also change over time, because he forgets some arguments.

We forcefully emphasize that we do not claim to be able to provide a complete account of all the elements encapsulated in this notion of perspective. In fact, our approach does not even require to believe that it is possible for anyone to capture the content of perspectives, or more generally to directly measure details about i's internal states of mind. The notion of perspective merely serves as an abstract device allowing to ground the idea that i may have changing attitudes towards some pairs of arguments.

Based on these notions, given T and S^*, define i's argumentative disposition towards T as $(\rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$. These three relations, described here below, constitute the formal primitives of our concept of argumentative disposition.

$\rightsquigarrow$ is a relation from S^* to T. An argument s *supports* a proposition t, denoted by $s \rightsquigarrow t$, iff i considers that s is an argument in favor of t. We emphasize that this definition should be understood in a conditional sense: $s \rightsquigarrow t$ means that i considers that, if s holds in her eyes, then she should endorse t, but this does not say anything about whether she thinks that s holds. An argument s may support several propositions in i's view, or none.

$\rhd_\exists$ is a binary relation over S^* representing whether i considers that a given argument trumps another one in *some* perspective. Let $s_1, s_2 \in S^*$ be two arguments. We note $s_2 \rhd_\exists s_1$ (s_2 *trumps* s_1) iff there is at least one perspective within which i considers that s_2 turns s_1 into an ineffective argument.[5] Let us emphasize that we are concerned with how i sees s_2

pirical, we propose to use another formalism, more adapted to our specific purpose, and leave aside here the task of more fully exploring the relations with proposals in formal argumentation theory such as dynamic argumentation.

[5]Note that, contrary to the usual assumption in formal argumentation theory, we do not consider it possible that both s_2 trumps s_1 and s_1 trumps s_2 *in a given perspective*. This is a choice of modelization, and not an hypothesis about the way i thinks: for $s_2 \rhd_\exists s_1$ to hold, by definition of our "trump" relation, s_2 must be a sufficiently strong

and s_1, not about whether s_2 should be considered to be a good argument to trump s_1 by any independent standard.

$\not\rhd_\exists$ is a binary relation over S^* defined in a similar way: $s_2 \not\rhd_\exists s_1$ iff there is at least one perspective within which i does not consider that s_2 turns s_1 into an ineffective argument.

We assume that $\forall s_2, s_1 \in S^* : \neg(s_2 \rhd_\exists s_1) \Rightarrow s_2 \not\rhd_\exists s_1$.

We consider that it is possible to query i about the trump relation between two arguments, and thus obtain information about $\rhd_\exists$, to the following limited extent: i may be presented with two arguments, s_1 and s_2, and asked whether he thinks that s_2 trumps s_1, or s_1 trumps s_2, or neither. In any case, we consider that i answers from the perspective he is currently in (to which we have no other access than through this query). Thus, if i answers that s_2 trumps s_1, we know that $s_2 \rhd_\exists s_1$. Indeed, in such a case we know that there is at least one perspective within which he thinks that s_2 trumps s_1: namely, the perspective that he currently has. Conversely, if i answers that s_2 does not trump s_1, we know that $s_2 \not\rhd_\exists s_1$.[6]

Remark 1. Whereas the two relations $(\rhd_\exists, \not\rhd_\exists)$ allow to capture i's changes of mind about whether a given argument can undermine another argument, the simple support relation $\rightsquigarrow$ adopted here does not permit to capture changes of mind about whether a given argument supports a given proposition. We assume that, in practice, when implementing our approach, propositions will be sufficiently simple and clear, so as to make it safe to assume that i will not change her mind concerning support during the decision process. This is a point to which the analyst will have to pay attention when applying our

argument to turn s_1 into an ineffective argument. If, on the contrary, i considers that s_2 is a plausible argument defending some claim incompatible with s_1, but not sufficiently strong to defeat s_1, then we model it by $s_2 \not\rhd_\exists s_1$ and $s_1 \not\rhd_\exists s_2$. Our choice permits to reduce our informational requirements, as there are fewer cases to be distinguished (our framework treats in the same way situations where two arguments trump each other and situations where none trumps the other). Note however that we do allow for the possibility that $s_2 \rhd_\exists s_1$ and $s_1 \rhd_\exists s_2$: this can happen by i adopting each of those two attitudes in two different perspectives. Hence, our choice of modelization does not translate in any formal restriction. This note only serves to make the semantics of the notion encapsulated by our "trump" relation clear.

[6] Another way of viewing the relations $\rhd_\exists$ and $\not\rhd_\exists$ goes as follows. Given a perspective p, define $\rhd_p$ as a binary relation over S^*: $s_2 \rhd_p s_1$ iff, when i is in the perspective p, s_2 turns s_1 into an invalid argument. Define P as the set of all possible perspectives. Then, define $\rhd_\exists = \bigcup_{p \in P} \rhd_p$, and $s_2 \not\rhd_\exists s_1$ iff $\exists p \in P \mid \neg(s_2 \rhd_p s_1)$. We favor another presentation because it emphasizes that we consider that we have direct access to $\rhd_\exists$ and $\not\rhd_\exists$, rather than to $\rhd_p$.

approach. If it appears, in real-life implementations, that this assumption is ill-advised, the framework will have to be extended by applying the approach used for $\rhd_\exists$ to the support relation (this would not raise any specific difficulty). For the time being, in the absence of empirical reasons to believe that the added generality is needed, we choose to use a single $\rightsquigarrow$ relation for simplicity. $\triangle$

Example 2 (Ranking (cont.)). Consider a set of criteria J. Consider the argument s_b = "Alternative a_1 ought to be ranked above a_2 because a_1 is better than a_2 on three criteria while a_2 is better than a_1 on only one criterion", and s_c = "It does not make sense to treat all criteria equally in this problem". Then (depending on i's disposition), it might hold that $s_c \rhd_\exists s_b$, and it might hold that $s_b \rightsquigarrow t_{a_1 \succ a_2}$. Note that both may very well hold together. $\triangle$

Definition 1 (Decision situation)**.** *We denote a* decision situation *by the tuple* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$*, with* $T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists$ *defined as above.*

The part of i's argumentative disposition that remains stable as i changes perspectives is of distinctive interest for decision analysis purposes. Indeed, recall that the emergence of new arguments may lead i to switch perspective. The stable part of her argumentative disposition is therefore a stance that proves resistant to the emergence of new arguments and is, in this sense, argumentatively well-grounded from i's point of view.

Let us therefore define the corresponding stable relations: $\rhd_\forall$ is defined as $s_2 \rhd_\forall s_1 \Leftrightarrow \neg(s_2 \not\rhd_\exists s_1)$. In plain words, $s_2 \rhd_\forall s_1$ if and only if there is no perspective within which s_2 does not trump s_1, or equivalently, $s_2 \rhd_\forall s_1$ if and only if s_2 trumps s_1 in all perspectives. Relatedly, $s_2 \not\rhd_\forall s_1$ is defined as: $s_2 \not\rhd_\forall s_1 \Leftrightarrow \neg(s_2 \rhd_\exists s_1)$. Hence, $s_2 \not\rhd_\forall s_1$ indicates that s_2 never trumps s_1. This implies, but is not equivalent to, $\neg(s_2 \rhd_\forall s_1)$.

Example 3 (Ranking (cont.)). Consider alternatives a_1 and a_2 such that a_1 Pareto-dominates a_2 on criteria J. Define s_d as an argument that states that a_1 ought to be ranked above a_2 because of the Pareto-dominance situation considering criteria in J. Then, it might hold that $s_d \rightsquigarrow t_{a_1 > a_2}$. Define s_f as "this is an incorrect reasoning because an important aspect to be considered in the problem is fairness and a_1 is worse than a_2 in this respect". Then it might be that $s_f \rhd_\exists s_d$ (assuming that i indeed considers fairness as important and that J does not include fairness). If i later changes her mind about the importance of fairness, then it will not hold that $s_f \rhd_\forall s_d$. $\triangle$

This enables us to define a decisive argument as one that is never trumped by any argument in S^*.

Definition 2 (Decisive argument)**.** *Given a decision situation* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$*, we say that an argument* $s \in S^*$ *is* decisive *iff* $\forall s' \in S^*$*:* $s' \not\rhd_\forall s$*.*

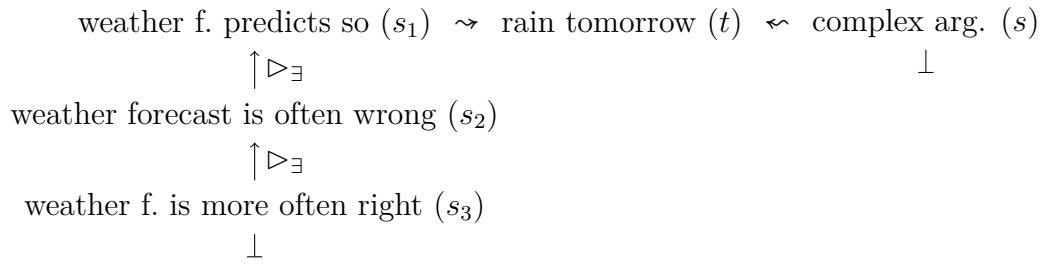

Figure 1: Illustration for examples 4 and 5. The symbol under s_3 and s indicates a decisive argument.

Notice that decisive arguments can be of very different sorts. Some decisive arguments will be very simple and straightforward arguments, which are so simple that they will be accepted by i whatever the perspective. By contrast, some decisive arguments will be very elaborate ones, taking many aspects of the topic into account and anticipating all sorts of arguments that could trump them, and accordingly never trumped by any other argument.

Example 4 (Weather forecast). Assume that individual i holds that t = "it will rain tomorrow" is supported by the argument s_1 = "one can expect that it will rain tomorrow because weather forecast predicts so". (See fig. 1.) But imagine that i also holds, at least from some perspective, that s_2 = "weather forecast is unreliable to infer what the weather will be like tomorrow because weather forecast is often wrong" is a counter-argument that trumps s_1. Imagine further that i would accept that an argument s_3 = "although it is often wrong, weather forecast is reliable because it is more often right than wrong" trumps s_2. Imagine, finally, that no argument trumps s_3 from any perspective.

In such a case, for i, s_1 is not a decisive argument. However, one can elaborate a more complex argument s = "weather forecast predicts that it will rain tomorrow. This may be an incorrect prediction, but weather forecast is more often right than wrong, thus its predictions constitute a sufficient basis to think that it will rain tomorrow". Notice that s includes the reasonings given by s_1 and s_3. Because s anticipates that s_2 could be envisaged to trump it, s could be decisive in supporting t (as assumed in fig. 1). $\triangle$

2.3. Deliberated judgment

Given a decision situation, we are now in a position to characterize i's stance towards the propositions in T once he has considered all the relevant arguments. We say that a proposition is justifiable if it is supported by a decisive

argument. A proposition is said to be untenable when each argument supporting it is always trumped by a decisive argument.

Definition 3 (Justifiable and untenable propositions)**.** *Given a decision situation* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$*, a proposition* t *is:*

- justifiable *iff* $\exists s \in S^* \mid s \rightsquigarrow t$ *and* $\forall s' : s' \not\rhd_\forall s$*;*
- untenable *iff* $\forall s \in S^* \mid s \rightsquigarrow t : \exists s_c \mid s_c \rhd_\forall s$ *and* $\forall s_{cc} : s_{cc} \not\rhd_\forall s_c$*.*

Three important aspects of this definition are worth emphasizing.

First, we use modal terms to name these notions: we talk about "justifi*able*" rather than "justifi*ed*" propositions. This is because, at a given point of time, individual i might well fail to accept, as a matter of brute empirical fact, a proposition supported by a decisive argument, for example, because she does not know this argument. Similarly, she might accept an untenable proposition. All this is despite the fact that the decisive arguments referred to in the definitions of justifiable and untenable propositions are decisive *according to* i*'s argumentative disposition* – that is, by i's own standards.

Second, notice that, according to our definition, a proposition can't be both justifiable and untenable, but it may be neither justifiable nor untenable. This may be the case if all the arguments supporting t have counter-arguments, but at least one argument supporting t has no decisive counter-argument.

Lastly, according to our definition, it is possible for a proposition t to be justifiable and for not-t, or more generally for any proposition t' in logical contradiction with t or having empirical incompatibilities with t, to be justifiable too. This specific definition allows to encompass situations in which there are intrinsically no more reason to accept t than t'. This can happen even when it is clear and evident for i that t and t' are incompatible, and even in situations where this incompatibility between t and t' is highlighted in some argument examined during the decision process.[7] This is a consequence of our definition of the trump relation, and it reflects the important idea that, as a matter of fact, in some decision situations, even if one takes all the relevant arguments into account, it can happen that several, mutually incompatible propositions are equally supported. It is part of the very aim of decision-aid, in such situ-

[7]Relatedly, notice that there is an important asymmetry between the notions of justifiable and untenable. Because t and some incompatible t' can both be justifiable, the fact that t is justifiable does not necessarily imply that the fate of t in i's view is entirely settled by its justifiability. By contrast, there is no way an untenable proposition could come back into the scene.

ations, to unveil the fact that mutually incompatible propositions are equally supported.[8]

Decision situations allowing to classify unambiguously all propositions in the agenda into justifiable or untenable propositions are of distinctive interest. Let us term such decision situations "clear-cut".

Definition 4 (Clear-cut situation)**.** *A decision situation* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$ *is* clear-cut *iff each proposition in* T *is either justifiable or untenable.*

Given a decision situation, we can now define i's DJ as those propositions $t \in T$ that are justifiable.

Definition 5 (DJ of i)**.** *The Deliberated Judgment corresponding to a decision situation* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$ *is:*

$$T_i = \{\, t \in T \;\mid\; t \text{ is justifiable} \,\}.$$

This notion of DJ, as we define it, captures what we take to be an important idea underlying Goodman's (1983) and Rawls' (1999) concept of "reflective equilibrium". This idea is that, if i manages, through an iterative process of revision of her opinion through the integration of new elements or arguments, to reach an "equilibrium" which is stable with respect to the integration of new elements, then the opinion reached at "equilibrium" is of distinctive interest – it captures i's "well-considered" or "true" opinion in some sense.[9]

Notice that the meaning of this definition depends on the interpretation given to S^* (see the beginning of section 2). In the idealistic interpretation, i's DJ is unique and fixed once and for all. In the pragmatic interpretation, i's DJ may evolve over time, as new arguments emerge.

Example 5 (Weather forecast (cont.)). To explain clearly this definition, it is useful to come back to our previous example (fig. 1) of individual i who holds that "weather forecast is often wrong" (s_2) is a counter-argument that trumps "it will rain tomorrow because weather forecast predicts so" (s_1). We have seen that a more complex argument (s), including both "weather forecast predicts that it will rain tomorrow" and an additional sub-argument that trumps s_2,

[8] Somewhat similar distinctions are discussed in formal argumentation theory about skeptical versus credulous justification (Prakken, 2006). Delving into the details of a comparative analysis falls beyond the scope of the present article.

[9] That said, our notion of DJ does not claim to reflect faithfully all the aspects of the notion of "reflective equilibrium" as used by the authors mentioned above. A thorough exploration of the links between our formal framework and these philosophical theories falls beyond the scope of the present article.

can turn out to be a decisive argument to support "it will rain tomorrow" (t). In such a case, t belongs to i's deliberated judgment, despite the fact that he might claim otherwise if not confronted with the complex argument above. △

Example 6 (Weather forecast (variant)). In this example T contains two propositions: t_1 is the proposition according to which it will rain tomorrow, and t_2 is the contrary proposition. Two corresponding arguments are s_1 and s_2 – two weather forecasts from different sources that predict respectively that it will rain and that it will not. Assuming that i attributes equal credibility to both sources and considers no other argument to be relevant, he might end up with both t_1 and t_2 in his deliberated judgment. This should not be interpreted as meaning that i is incoherent, but simply as a situation where different propositions are equally justified for lack of means to tell them apart. Similarly, scientists can consider two contradictory hypotheses plausible, for lack of current knowledge; or someone may hold that two incompatible acts are equally (im)moral. △

3. Issues of empirical validation

The former section clarified definitions and explained the articulations between the key concepts of our framework, at a rather abstract level. Now we want to investigate how this framework can be confronted with empirical reality. For that purpose, we will examine how one can test a *model* of the support and trump relations built by a decision analyst trying to capture the deliberated judgment of a decision-maker.

Let us define a model η of a decision situation as a pair of relations $\rightsquigarrow_\eta \subseteq S^* \times T$ and $\rhd_\eta \subseteq S^* \times S^*$. These relations are not necessarily an approximation of the real $\rightsquigarrow, \rhd_\exists$ relations characterizing i. Indeed, the chief aim of the model is to know i's DJ, not to reflect in detail what i thinks about all arguments, which would arguably not be achievable (we will come back to this important point below).

Define T_η as the set of propositions that the model η claims are supported:

$$T_\eta = \rightsquigarrow_\eta(S^*) = \{\, t \in T \;\mid\; \exists s \in S^* \mid s \rightsquigarrow_\eta t \,\}.$$

Example 7 (Ranking (cont.)). We have already defined a set of alternatives $\mathcal{A}$, propositions T representing possible comparisons of the alternatives, and criteria J. Consider further a set of criteria functions $(g_j)_{j\in J}$ evaluating all the alternatives $a \in \mathcal{A}$ using real numbers: $g_j : \mathcal{A} \to \mathbb{R}$.

Imagine that i's problem is to decide which kind of vegetable to grow in his backyard. Assume an analyst providing decision-aid to i considers that the problem can be reduced to a ranking between three candidates: carrots, lettuce and pumpkins, denoted by $c, l, p \in \mathcal{A}$. The analyst believes that i is ready to rank vegetables according to exactly two criteria. The analyst has obtained six real numbers $g_j(a)$, representing the performances of each alternative on each criteria, and believes that i is ready to rank vegetables according to the sum of their performances on the two criteria, $v(a) = g_1(a) + g_2(a)$.

The analyst can now try to represent i's attitude using a model $\eta = (\rightsquigarrow_\eta, \rhd_\eta)$ by producing sentences that explain to i the "reasoning" underlying the definition of v. Assume the values given by v position carrots as winners. The analyst could define an argument $s_{(c,l)}$ "carrots are a better choice than lettuce because carrots score $g_1(c)$ on criterion one, and $g_2(c)$ on criterion two, which gives it a value $v(c)$, whereas lettuce scores $g_1(l)$ on criterion one, and $g_2(l)$ on criterion two, which gives it an inferior value $v(l)$". In the model of the analyst, this argument supports the proposition that carrots are ranked higher than lettuce: $s_{(c,l)} \rightsquigarrow_\eta t_{c \succ l}$. The model contains similar arguments in favor of other propositions $t \in T$ that are in agreement with the values given by v. In our example, the analyst furthermore believes that no counter-arguments are necessary and thus defines $\rhd_\eta = \emptyset$. △

3.1. Validity and the problem of observability

Because the point of carving out η is to capture i's Deliberated Judgment T_i, we can define a valid model as one that correctly captures T_i.

Definition 6 (Validity)**.** *A model η is valid iff $T_\eta = T_i$.*

How can the analyst determine if a given model η is a valid one?

Let us assume that the only information that he can use for that purpose is the one he can get by querying i – and is, in that sense, "observable" for him. DJs are not observable in that sense. Indeed, i's DJ are defined in terms of $\not\rhd_\forall$. But observing $\not\rhd_\forall$ would require that i takes successively all the possible perspectives she can have, which is unrealistic.[10]

In the remainder of this section, we explain how we handle this conundrum in two steps. First, section 3.2 introduces a provisional solution, by identifying conditions that guarantee the existence of a model allowing to identify i's DJ on the basis of what we will call an "operational" validity criterion – that is,

[10] This would amount to assume that i already knows all the arguments and can aggregate them successfully. If this were possible, i would probably not need help from an analyst.

a criterion based on observable data. Then, section 3.3 explores how these conditions can be weakened.

3.2. Existence of a valid model and its conditions

In this subsection, we introduce apparently reasonable conditions about the way i reasons and about the decision situation. Our theorem will then guarantee that a model exists and captures correctly i's DJ if those conditions are satisfied on S^* and if the model satisfies a validity criterion that, as opposed to validity itself, can be directly checked on the basis of observable data (an "operational validity" criterion).

3.2.1. Conditions

A first condition about $\rhd_\exists$ mandates a certain form of stability. It assumes that i possibly changes her mind about whether an argument s' trumps another one only when there exists another argument that trumps s'.

Condition 1 (Answerability)**.** *A decision situation* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$ *satisfies* Answerability *iff, for all pairs of arguments* (s, s')*:*

$$s' \rhd_\exists s \text{ and } s' \not\rhd_\exists s \Rightarrow \exists s_c \mid s_c \rhd_\exists s'.$$

Let us now turn to the second condition. It has to do with the way i reasons. Imagine that i finds himself in the following uneasy situation. He declares that s_1 is trumped by s_2. However, i is also ready to declare that s_2 is in turn trumped by s_3, a decisive argument. In such a situation, it seems natural enough to assume that, if we carve out an argument s, playing the same argumentative role as s_1, but anticipating and defeating attempts to trump it using s_2, i will endorse s.

This assumption is formalized by the condition *Closed under reinstatement* below. To write it down, we first need to formalize, thanks to the following notion of *replacement*, the idea that a set of arguments is at least as powerful as another argument, from the point of view of its argumentative role. We say a set of arguments $S \subseteq S^*$ replaces an argument $s \in S^*$ whenever all the arguments trumped by s are also trumped by some argument $s' \in S$, and all the propositions supported by s are also supported by some argument $s' \in S$.[11]

[11] Note that the replacer may be more powerful than the argument it replaces, in the sense that it may trump arguments or support propositions than the replaced argument did not trump or support.

Definition 7 (Replacing arguments)**.** *A set of arguments* $S \subseteq S^*$ replaces $s \in S^*$ *iff* $\rhd_\exists(s) \subseteq \rhd_\exists(S)$ *and* $\rightsquigarrow(s) \subseteq \rightsquigarrow(S)$*. We say that* s' *replaces* s*, with* $s, s' \in S^*$*, to mean that* $\{s'\}$ *replaces* s*.*

Condition 2 (Closed under reinstatement)**.** *A decision situation* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$ *is* closed under reinstatement *iff,* $\forall s_1 \neq s_2 \neq s_3 \neq s_1 \in S^*$ *such that* $s_3 \rhd_\forall s_2 \rhd_\exists s_1$*, with* s_3 *decisive:*

$$\exists s \mid s \text{ replaces } s_1 \text{ and } \rhd_\exists^{-1}(s) \subseteq \rhd_\exists^{-1}(s_1) \setminus \{s_2\}.$$

The condition mandates that, whenever some decisive argument always trumps s_2, which in turn trumps s_1, it is possible to replace s_1 by an argument that is no longer trumped by s_2 and is not trumped by any other argument than those trumping s_1.[12]

Finally, we introduce two conditions on the size of the relation $\rhd_\exists$.

Let us call a chain of length k in $\rhd_\exists$ a finite sequence s_i of arguments in S^*, $1 \leq i \leq k$, such that $s_i \rhd_\exists s_{i+1}$ for $1 \leq i \leq k-1$. An infinite chain is an infinite sequence s_i such that $s_i \rhd_\exists s_{i+1}$ for all $i \in \mathbb{N}$.

Condition 3 (Bounded width)**.** *A decision situation* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$ *has a* bounded width *iff there is no argument that is trumped by an infinite number of counter-arguments.*

Condition 4 (Bounded length)**.** *A decision situation* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$ *has a* bounded length *iff there is no infinite chain in* $\rhd_\exists$*. (Cycles in* $\rhd_\exists$ *are therefore excluded as well.)*

3.2.2. Operational validity criterion

Let us now define the following "operational" validity criterion for a model η intended to capture i's DJ. We term it "operational" to emphasize that, as opposed to the definition of validity (definition 6), it can be checked on the sole basis of observable data.

Definition 8 (Operational validity criterion)**.** *A model* η *of a decision situation is operationally valid iff, whenever* $(s \rightsquigarrow_\eta t)$*, it holds that* $[s \rightsquigarrow t]$ *and* $[\forall s_c \in S^* : (s_c \not\rhd_\exists s) \vee (\exists s_{cc} \rhd_\eta s_c \wedge s_{cc} \rhd_\exists s_c)]$*, and whenever* t *is not supported by* η*,* $\forall s \rightsquigarrow t : \exists s_c \rhd_\eta s \wedge s_c \rhd_\exists s$*.*

[12] Such a configuration of arguments, where s_3 trumps s_2 which in turns trumps s_1, recalls the notion of "strong defense" in argumentation theory (Baroni and Giacomin, 2007). A further discussion of this issue falls beyond the scope of this paper.

This criterion amounts to partially comparing, on the one hand, i's argumentative disposition towards propositions and arguments and, on the other hand, η's representations of i's argumentative disposition.[13] More precisely, a model satisfies the operational validity criterion (for short: is operationally valid) iff:

(i) arguments that, according to the model, support a proposition t are indeed considered by i to support t;

(ii) whenever a model uses an argument s to support a proposition, and that argument is trumped by a counter-argument s_c, the model can answer with a counter-counter-argument, using a counter-counter-argument that i confirms indeed trumps the counter-argument s_c;

(iii) whenever an argument s supports a proposition that the model does not consider to be supported, the model is able to counter that argument using a counter-argument that i confirms indeed trumps s.

As required, this criterion is uniquely based on observable data. Indeed, recall that the only observable data that the analyst can use are the ones obtained by querying i by asking her if a given argument s_2 trumps another argument s_1. If she replies that it does, this is enough to conclude that, according to her, $s_2 \rhd_\exists s_1$. Indeed, in such a case, we know that there is at least one perspective within which she thinks that s_2 trumps s_1: namely, the perspective that she currently has. Querying i can thus provide the information needed to check if a model is operationaly valid.

3.2.3. Theorem

Because querying i will not give enough information to know that $s_2 \rhd_\forall s_1$ (if indeed $s_2 \rhd_\forall s_1$), querying i will never allow to directly claim that a model satisfies the definition of validity (definition 6). What we need therefore is a means to ensure that an operationally valid model is a valid one. This is provided by the following theorem.

Theorem 1. *Assume a decision situation $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$ is Closed under reinstatement, Answerable and has Bounded length and width. Then: i) the decision situation is clear-cut; ii) there exists an operationally valid model of that decision situation; iii) any operationally valid model η satisfies $T_i = T_\eta$.*

[13] This procedure could be considered as a persuasion dialogue (Prakken, 2009).

Theorem 2 (in section 3.3) generalizes this theorem. It is proven in appendix A.

Example 8 (Budget reform). Let us take a non trivial example that will be used to illustrate how theorem 1 can be used and why we need to go beyond this first theorem. Imagine that i is a political decision-maker. She wants to run for an election, and is elaborating her policy agenda. She has heard about Meinard et al.'s (2017) (thereafter referred to as "M") argument that, according to a popular survey, biodiversity should be ranked after retirement schemes and public transportation, but before relations with foreign countries, order and security, and culture and leisure in the expenses of the State. Assume that i wants to make up her mind about the single proposition $t =$ "I should include in my agenda a reform to increase public spending on biodiversity conservation so as to rank biodiversity higher than relations with foreign countries in the State budget".

She requests the help of a decision analyst. The latter starts by reviewing the literature to identify a set of arguments with which he will work. (The arguments are illustrated in fig. 2.) He thereby identifies that proposition t can be considered to be supported by $s =$ "M's finding (stated above) is based on a large scale survey and quantitative statistical analysis, and their protocol was designed to track the preferences that citizens express in popular votes. There are therefore scientific reasons to think that a policy package including the corresponding reform will gather support among voters." Pursuing his exploration of the recent economic literature on environmental valuation methods, the analyst could identify only two counter-arguments to s:

- $s_{c1} =$ "M's measure is extremely rough as compared to more classical economic valuations, such as contingent valuations and the like (Kontoleon et al., 2007), which makes it non credible as a guide for policy";
- $s_{c2} =$ "M claim to value biodiversity *per se*. The very meaning of such an endeavor is questionable because it is too abstract. More classical economic valuations are focused on concrete objects and projects, which is more promising".

But he also found a counter-counter-argument to each of these counter-arguments:

- $s_{c1c} =$ "Biodiversity is not the kind of thing about which people make decisions in their everyday life. Their preferences about it are accordingly likely to be rough. The exceedingly precise measurements provided by contingent valuations and the like are therefore more a weakness than a strength";

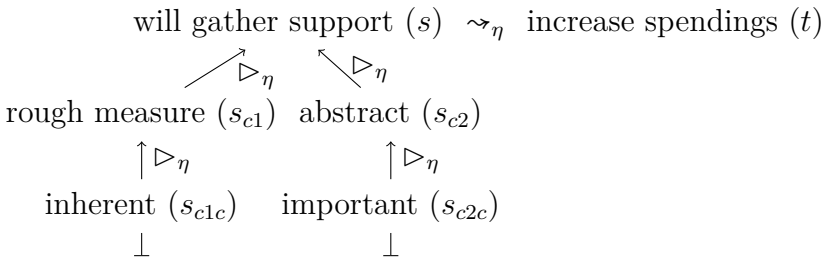

Figure 2: Illustration for example 8. (Only the arguments used by the model η are displayed.)

- s_{c2c} = "Abstract notions such as biodiversity are an important determining factor for many people when they make decisions. Eschewing to value them is ill-founded".

Imagine further that the analyst has not found any argument liable to trump either s_{c1c} or s_{c2c}.

Define $s_{1,\text{reinstated}}$ as: "[*content of* s]; this is a rough measure but [*content of* s_{c1c}]"; similarly, define $s_{2,\text{reinstated}}$ as "[*content of* s]; the very meaning could be questioned because it is highly abstract, but [*content of* s_{c2c}]"; and define $s_{\text{reinstated}}$ as "[*content of* s]; this is a rough measure but [*content of* s_{c1c}]; the very meaning could be questioned because it is highly abstract, but [*content of* s_{c2c}]". Define $S \subseteq S^*$ as the set of argument comprising s, s_{c1}, s_{c2}, s_{c1c}, s_{c2c}, $s_{1,\text{reinstated}}$, $s_{2,\text{reinstated}}$ and $s_{\text{reinstated}}$.

Assume that the analyst is justified to think that i's reasoning is such that S^* satisfies Closed under reinstatement, Answerability, Bounded length and Bounded width. Recall now that, in order to identify the propositions lying in T_i, the analyst must identify arguments supporting propositions in T_i, such that these arguments can resist counter-arguments from the whole of S^*. In other words, the analyst must test the claims of the model not only against the counter-arguments in S, but against the whole of S^*, which the analyst ignores.

Imagine now that the analyst assumes that, even though S is a strict subset of S^*, S is a good enough approximation of S^*, in the sense that there is no argument in $S^* \setminus S$ that trumps any argument in S or that supports t. Thanks to theorem 1, the analyst can then deduce that the situation is clear-cut and that there exists a valid model of the decision situation.

The next step for him is to carve out a model η reproducing the relations between arguments that he found in the literature, and then to test whether his model is operationally valid using definition 8. In order to validate η, he

would first ask i whether she agrees that s supports t. If so, he then would check whether i considers that s_{c1} is a counter-argument to s, in which case the analyst would check that the counter-counter-argument that he envisaged, s_{c1c}, is considered by i to trump s_{c1}. The analyst would then proceed in a similar way with the second chain of counter-arguments (s_{c2} and s_{c2c}), and verify that, as η hypothesizes, i does not take any other argument in S to trump s. This would, eventually, allow him to conclude on the validity of the model η. Should it prove operationally valid, the analyst could then conclude that $T_i = \{t\}$ (using theorem 1 and $T_\eta = \{t\}$).

But notice that this whole story only works because we assumed that arguments in $S^* \setminus S$ never trump any argument in S. This assumption is clearly unrealistic: any slight reformulation of s_{c1}, for example, will most likely also trump s. This is not the only unrealistic assumption in our hypothetical scenario: it is also unlikely that the whole set S^* indeed satisfies Bounded length, for example. This condition requires an absence of cycle in the trump relation. While this may be considered to hold on S, it is possible that some ambiguous or poorly phrased arguments in S^* would confuse i in such a way that i will declare, for example, that $s_1 \rhd_\exists s_2 \rhd_\exists s_3 \rhd_\exists s_1$ for some triple of such unclear arguments. Hence the need to go beyond theorem 1. $\triangle$

Theorem 1 embodies an important step towards being able to confront models of deliberated judgment with empirical reality, by spelling out sufficient conditions upon which unrolling the procedures of refutation is not a pure waste of time and energy, because there is something to be found. It also illustrates the potential usefulness of the notion of operational validity. Indeed, since the point of the modeling endeavor in our context is to capture T_i, we know by virtue of iii) in theorem 1 that, if the corresponding conditions are met, and if we have good reasons to believe that we have an operationally valid model, then we can admit that it captures T_i.

However, establishing this theorem cannot be more than just a first step. As illustrated in example 8, the conditions above are quite heroic. One cannot realistically expect that real-life decision situations will fulfill these conditions. The most important issue is that we need a means to distinguish S^* from the restricted set of arguments with which the analyst works in practice. And we need means to make sure that the restricted set indeed "covers" the matter "sufficiently", so as to escape the situation in which the analyst is locked in example 8, where he finds himself condemned to make wildly unrealistic assumptions. The next subsection tackles this pivotal issue.

3.3. Weakening of some conditions

To obtain the results we want, all we actually need is that it should be possible to define a subset of arguments $S_\gamma \subseteq S^*$ that satisfies conditions akin to the ones defined above, and which are sufficient to cover the topic at hand.

Let us start by formalizing the requirement, for S_γ, to cover the topic at hand. What we want is that all the arguments needed for the decision-maker to make up her mind about the topic should be encapsulated in S_γ. This means that, if arguments from $s \in S^* \setminus S_\gamma$ are brought to bear, it should be possible either to discard them or to show that they can be replaced by arguments in S_γ.

This is done thanks to the following formal definitions and condition.

Definition 9 (Unnecessary argument)**.** *Given a decision situation and a subset $S_\gamma \subseteq S^*$ of arguments, we say that $S \subseteq S^*$* essentially replaces *$s \in S^*$ iff $(\rhd_\exists(s) \cap S_\gamma) \subseteq \rhd_\exists(S)$ and $\rightsquigarrow(s) \subseteq \rightsquigarrow(S)$.*

Let $S_{\gamma\text{dec}} = S_\gamma \cap \overline{\rhd_\exists(S^*)}$ denote the decisive arguments in S_γ. We say that an argument $s \in S^*$ is resistant *iff it is not trumped by any argument in $S_{\gamma\text{dec}}$. Let $S_{\gamma\text{res}} = S_\gamma \cap \overline{\rhd_\exists(S_{\gamma\text{dec}})}$ denote the resistant arguments in S_γ.*

We say that an argument $s \in S^*$ is unnecessary *iff s is trumped by a resistant argument from S_γ or s is essentially replaceable by $S_{\gamma\text{res}}$. In formal terms: $s \in \rhd_\exists(S_{\gamma\text{res}})$ or $[(\rhd_\exists(s) \cap S_\gamma) \subseteq \rhd_\exists(S_{\gamma\text{res}})$ and $\rightsquigarrow(s) \subseteq \rightsquigarrow(S_{\gamma\text{res}})]$.*

Condition 5 (Covering set of arguments)**.** *Given a decision situation and a set of arguments $S_\gamma \subseteq S^*$, S_γ is* covering *iff all arguments $s \in S^* \setminus S_\gamma$ are unnecessary.*

Let us now relax the conditions of theorem 1 by formulating weaker requirements confined to S_γ. This adaptation is straightforward for conditions 1 and 2.

Condition 6 (Set of arguments allowing answerability)**.** *Given a decision situation and a subset $S_\gamma \subseteq S^*$ of arguments, we say that the set S_γ satisfies* Answerability *iff, for all $s \in S^*, s' \in S_\gamma$: $s' \rhd_\exists s$ and $s' \not\rhd_\exists s \Rightarrow \exists s_c \in S^* \mid s_c \rhd_\exists s'$.*

Condition 7 (Set of arguments closed under reinstatement)**.** *Given a decision situation $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$ and a subset $S_\gamma \subseteq S^*$ of arguments, we say that the set S_γ is* closed under reinstatement *iff, $\forall s_1, s_3 \in S_\gamma, s_1 \neq s_3, s_3$ not trumping s_1, s_3 decisive:*

$$\exists s \in S_\gamma \mid s \text{ replaces } s_1 \text{ and } \rhd_\exists^{-1}(s) \subseteq \rhd_\exists^{-1}(s_1) \setminus \rhd_\forall(s_3).$$

This condition is vacuous when there is no s_2 such that $s_3 \rhd_\exists s_2 \rhd_\exists s_1$: in that case, s_1 replaces itself.

Similarly, we can relax condition 3 and apply it to a subset of arguments. When an argument has very numerous counter-arguments, one may think that their vast number might spring from some common reasoning that they share. For example, an argument might involve some real value as part of its reasoning, and be multiplied as infinitely many similar arguments of the same kind using tiny variations of that real value. If so, and if we know that we can convincingly rebut each of these counter-arguments, we might believe that only a small number of counter-counter-arguments will suffice to rebut the counter-arguments.

Definition 10 (Defense)**.** *We say $s \in S^*$ is S_γ-*defended *iff all the arguments s_c trumping s are trumped by a decisive argument in S_γ, or formally, $\forall s_c \in S^* \mid s_c \rhd_\exists s : (\exists s_{cc} \in S_\gamma \mid s_{cc} \rhd_\forall s_c, s_{cc}$ decisive$)$. We say $s \in S^*$ is (j, S_γ)-*defended *iff there exists a set $S \subseteq S_\gamma$ of arguments of cardinality at most j such that s is S-defended (thus, if j arguments from S_γ suffice to defend s).*

Condition 8 (Set of arguments with width bounded by j)**.** *Given a decision situation and a natural number j, a set of arguments $S_\gamma \subseteq S^*$ has* width bounded by j *iff, for each argument $s \in S_\gamma$, if s is S_γ-*defended*, then it is (j, S_γ)-defended.*

The condition is vacuously true when no argument in S^* is trumped by more than j counter-arguments.

Our last condition relaxes condition 4. We want to exclude *some* of the long chains in S^*. But we want to tolerate long chains, including cycles, among unclear arguments. Indeed, anecdotal evidence from ordinary argumentation situations suggests that in many (otherwise interesting) decision situations, cycles do appear in trump relations among arguments (for example, because arguments can use ambiguous terms). However, this does not necessarily prevent the situation from being modelizable in our sense. What we do need is to avoid some of the cycles or chains that involve "too many" arguments from S_γ, in a somewhat technical sense captured by the following condition.

Condition 9 (Set of arguments with length bounded by k)**.** *Given a decision situation, a natural number k, and a set of arguments S_γ, define a binary relation Q over S_γ as $s_2 Q s_1$ iff $s_2 \rhd_\exists s_1$ or $s_2 \rhd_\exists s \rhd_\exists s_1$ for some $s \in S^*$, thus, $Q = (\rhd_\exists \cup (\rhd_\exists \circ \rhd_\exists)) \cap (S_\gamma \times S_\gamma)$. Let $Q^1 = Q$ and $Q^{k+1} = Q^k \circ Q$ for any natural number k. The set S_γ has length bounded by k iff $\nexists s_2, s_1 \in S_\gamma \mid s_2 Q^{k+1} s_1$, thus, iff it is impossible to reach an argument from S_γ, starting from an argument from S_γ, following Q more than k times.*

This condition tolerates cycles[14] in $\rhd_\exists$ that involve only arguments picked outside the chosen set S_γ. It only forbids a subset of the situations where a cycle (or a too long chain) is built that involve arguments from S_γ. For example, it excludes a situation where $s_2 \rhd_\exists s \rhd_\exists s_1 \rhd_\exists s_2$ for some $s_1, s_2 \in S_\gamma$ and $s \notin S_\gamma$.[15]

Thanks to conditions 5 to 9, we are now in a position to define our set of arguments of interest.

Definition 11 (CAC arguments)**.** *Given a decision situation and a set $S_\gamma \subseteq S^*$, we say that S_γ is* clear *and* covering*, or CAC, iff it is Closed under reinstatement and Answerable, and has width bounded by some number j and length bounded by some number k, and is such that all arguments $s \in S^* \setminus S_\gamma$ are unnecessary.*

Following the same rationale, we can define an operational criterion echoing definition 8.

Definition 12 (S_γ-operational validity)**.** *Given a decision situation and a set $S_\gamma \subseteq S^*$, we define a model η as S_γ-*operationally valid *iff for all $(s \rightsquigarrow_\eta t)$, $s \in S^*$, we have $[s \rightsquigarrow t]$ and $[\forall s_c \in S_\gamma : (s_c \not\rhd_\exists s) \vee (\exists s_{cc} \in S^* \mid s_{cc} \rhd_\eta s_c \wedge s_{cc} \rhd_\exists s_c)]$, and when t is not supported by η, $\forall s \in S_\gamma \mid s \rightsquigarrow t : (\exists s_c \in S^* \mid s_c \rhd_\eta s \wedge s_c \rhd_\exists s)$.*

A theorem echoing theorem 1 can then be proved.

[14] Cycles in our sense have to be distinguished from cycles involving an attack relation as defined in formal argumentation theory. We do not deny that cycles of attacks in the formal argumentation sense often happen, and condition 9 does not exclude cycles understood in that sense: these cycles are generally not cycles in "trump" relations. We consider that an argument s_2 trumps another one only when i considers that the first one is strong enough to render the second one ineffective. This definition relies on an asymmetry, s_2 being, in a sense, "favored over" s_1. Our trump relation is therefore somewhat analogical to a strict preference relation, for which an assumption of acyclicity is commonplace in the literature.

[15] Readers used to decision theoretic axiomatizations might find this condition odd, since axioms usually mandate conditions considered more "basic", such as transitivity and irreflexivity, and derive from them the conclusion that cycles are forbidden. This strategy does not work for our setting (or is not applicable in a simple way), because "basic" conditions such as transitivity would be unreasonable to impose here. For example, given $s_3 \rhd_\exists s_2$ and $s_2 \rhd_\exists s_1$, it is easy to think about situations where i would consider that $s_3 \not\rhd_\forall s_1$, and to think about situations where i would consider that $s_3 \rhd_\exists s_1$. Neither anti-transitivity nor transitivity can thus be reasonably imposed (and our current condition avoids such requirements). Studying which conditions exactly are necessary to ban cycles (or make them innocuous) in our setting would be interesting, but it does not seem crucial at this stage. Indeed, in concrete settings we consider that cycles involving arguments from S_γ are unlikely to occur. (This claim should be backed up by empirical studies.)

Theorem 2. *Given a decision situation* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$*, given* $S_\gamma \subseteq S^*$*, if* S_γ *is CAC, then i) the decision situation is clear-cut; ii) there exists an* S_γ*-operationally valid model* η*; iii) any* S_γ*-operationally valid model* η *satisfies* $T_i = T_\eta$*.*

This theorem is a strengthened version of theorem 1 since it produces the same results based i) on the conditions encapsulated in the definition of CAC arguments, and ii) on S_γ-operational validity. Those conditions are implied by the ones assumed by theorem 1. Indeed, when the conditions of theorem 1 hold, taking $S_\gamma = S^*$ satisfies the conditions of theorem 2.[16]

4. Significance of the deliberated judgment framework for decision theory and the practice of decision analysis

Section 2 displayed the conceptual core of our framework and section 3 explained how this framework can be confronted to empirical reality. The present section reflects on the meaning, promises and limits of our approach. We start by pondering on how the various conditions spelled out in section 3 can be interpreted (section 4.1). We then take a broader view to discuss how our framework relates to the larger literature in decision science (section 4.2).

4.1. The meaning of our conditions

In order to understand the precise meaning of the conditions of theorem 1 and, more importantly, of theorem 2, an almost trivial but nonetheless very important first step is to spell out what it means if these conditions are *not* fulfilled.

We already stressed that the conditions of theorem 1 are certainly too strong to be fulfilled. The conditions of theorem 2 are, by construction, much weaker.

[16] Theorem 2 has an interesting corollary which permits to view our proposal as providing useful means to take account of the fact that knowledge evolves. In some cases it might be important, for example for efficiency reasons in contexts of limited resources, to investigate if a decision-aid provided before some discovery of new knowledge is still valid after the discovery. Take a decison-aid which has been provided using a set of argument S_γ which is CAC with respect to the set of known arguments before the discovery S^*_{before} and using a S_γ-operationally valid model η. Theorem 2 shows that, if we can prove that S_γ is CAC with respect to the set of all the arguments S^*_{after} supplemented thanks to the new discovery, then there is no need to check the validity of η again.

But still, there certainly are situations where they are not fulfilled. In such cases, we do not claim that decision analysis is impossible. Neither is our general framework, as presented in section 2, rendered bogus. The sole implication is that our approach to operational empirical validation cannot be implemented. This does not prevent, for example, the analyst from trying to identify directly decisive arguments, and this does not render irrelevant a decision analysis based on decisive arguments. Neither does this prevent completely other approaches to decision analysis to be implemented. The only implication is that a full-fledged implementation of our approach, including operational empirical validation, is not guaranteed to be possible in such situations. It is no part of our claim that our approach can be applied all the time and provides an all-encompassing framework liable to overcome all other approaches to decision analysis. Our approach has a specific domain of application.

Beyond these simple, negative comments, how are our conditions to be understood? In general terms, these various conditions can be interpreted in three different ways:

(i) as axioms capturing minimal properties concerning arguments and the way i reasons,

(ii) as empirical hypotheses,

(iii) as rules governing the decision process (rules that i can commit to abide by, or can consider to be well-founded safeguards for the proper unfolding of the process).

Example 9 (Budget reform (cont.)). We can now improve example 8 by relaxing the assumptions it contains. One can envisage in turn the three possibilities spelled out above.

In interpretation (i), instead of assuming that i always reasons in such a way that S^* in its entirety satisfies the conditions of theorem 1, we only assume that the set of argument $S = \{s,\ s_{c1},\ s_{c2},\ s_{c1c},\ s_{c2c},\ s_{1,\text{reinstated}},\ s_{2,\text{reinstated}},\ s_{\text{reinstated}}\}$ is CAC.

In interpretation (ii), we have to take advantage of empirical data to claim that the above set is CAC. Imagine, for example, that we have been able to show that the overwhelming majority of people does reason with respect to the arguments in this set in such a way that it can be considered CAC. This would provide strong empirical support to admit that this set can be considered CAC for the purpose of the decision process at issue (assuming the pragmatic interpretation of S^*). In the present article, we leave aside the important difficulties that such empirical concrete applications would face.

In interpretation (iii), the analyst would start by explaining to i the content of the requirements encapsulated in the definition of a CAC set of arguments and ask her if she is willing to commit herself to reason in such a way as to fullfill these requirements when thinking about the arguments to be discussed in the process. For example, for the Answerability of the set of arguments (condition 6), the analyst would ask i if she would accept to commit not to change her mind depending on her mood or any other non-argumentative factor. Notice that i might figure at some point that it was not a good idea after all to commit to these various things, and in such a case the decision analysis process would fail. △

Some of the conditions of our theorems are arguably more congenial to a given interpretation. For example, it seems natural enough to interpret condition 2 as a rationality requirement of the kind that it makes sense to use as an axiom (interpretation (i)). By contrast, condition 1 is the kind of condition that can easily be translated in the form of rules than decision-makers can be asked to abide by when they engage in a decision process (interpretation (iii)). By construction, conditions 6 and 7 are weakened versions of the above stronger conditions. They accordingly inherit the preferred interpretation suggested above. Conditions 8 and 9 can easily be seen as empirical hypotheses (interpretation (ii)).

However, although it is tempting to draw such connections between specific conditions and specific interpretations, at a more abstract level all the conditions above can be interpreted in all three interpretations. The different conditions can even be interpreted differently in the context of different implementations. In the present, largely theoretical work, we want to leave all these possibilities open. Future, more applied works, should assess if and when these different interpretations can be used, in particular by elaborating and implementing the convenient empirical validation protocols in interpretation (ii) and the convenient participatory procedures in interpretation (iii).

4.2. The deliberated judgment framework in perspective

Now that the meaning of the conditions of our theorems is clarified, we are in a firmer position to discuss the nature of our contribution to the literature.

The central, distinctive concept of our approach is the one of deliberated judgments of an individual. Deliberated judgments are the propositions that the individual herself considers based on decisive arguments, on due consideration. This formulation highlights the two key features of the concept.

The first key feature is that deliberated judgments are the result of a careful

examination of arguments and counter-arguments. This echoes the approach to the notion of rationality developed most prominently by Habermas (1981). In this approach, actions, attitudes or utterances can be termed "rational" so long as the actor(s) performing or having them can account for them, explain them and use arguments and counter-arguments to withstand criticisms that other people could raise against them. Variants of this vision of rationality play a key role in other prominent philosophical frameworks, such as Scanlon's (2000) and Sen's (2009). Having in mind this approach to rationality, in the remainder of this discussion, we will therefore simply talk about "rationality" when referring to this first idea underlying our framework.

The second key feature is that deliberated judgments are nevertheless the individual's own judgments, in the sense that they do not reflect the application of any exogenous criterion. This second idea can also be nicknamed, for brevity's stake, by simply talking about "non-paternalism".

Our approach, when applied in a decision analysis perspective, requires admitting the soundness of these two normative notions of rationality and non-paternalism.

Our approach however also has a strong descriptive dimension, which is a direct implication of the very meaning of non-paternalism. Though we are interested in deliberated judgments rather than in the "shallow" preferences that individual spontaneously express, still the deliberated judgments that we are interested in are the ones of real, empirical individuals that are not constrained by our framework to adhere to a specific set of exogeneous stances. These descriptive aspects feed a normative approach that accordingly owes its normative credentials both to its normative foundations and to its reference to empirical reality.

Due to this double anchorage in normative and descriptive aspects, our approach opens avenues to overcome perennial difficulties facing decision theory concerning its descriptive vs. normative status. Indeed, our framework sets the stage for decision-aiding practices that could have a crucial strength as compared with more standard approaches, by including rigorous tests of whether individuals endorse or not various arguments and argumentative lines, thereby avoiding both actively advocating them (a purely normative approach) and leaving the individual in the ignorance of their existence (a purely descriptive approach). Decision analyses based on deliberated judgments thereby provide compelling reasons for the aided individual to think that the decisions he makes once he has been aided are better than the one he would have made otherwise. Such reasons are liable to play a key role in strengthening the legitimacy and validity of decision analysis – two requirements largely discussed in the literature (Landry et al., 1983, 1996).

In order to illustrate this idea, it is useful to compare our framework to more classical approches, such as utility theory. Proponents of utility theory could claim that utility functions provide arguments that individuals will consider convincing (Savage, 1972; Morgenstern, 1979; Raiffa, 1985), and that therefore our approach will converge towards utility theory. However, the convincing power of utility-based arguments is debatable (Ellsberg, 1961; Allais, 1979). Psychologists have tried to test it experimentally (Slovic and Tversky, 1974; MacCrimmon and Larsson, 1979). But such tests can hardly be considered conclusive: the meaning of their results depends on how arguments have been presented to the individuals and on whether counter-arguments have been presented, as Slovic and Tversky (1974) themselves point out. Such a systematic confrontation with counter-arguments is precisely what our proposed framework allows to implement.

The formal framework presented in this article will however only live up to its promises if empirical applications are developed. Researchers in artificial intelligence (Labreuche, 2011) and persuasion (Carenini and Moore, 2006) have produced ways of "translating" formal Multi-Attribute Value Theory models into textual arguments, that could possibly provide promising tools to develop such applications.

Acknowledgements

We thank Denis Bouyssou, Cyril Hédoin, Jean-Sébastien Gharbi, André Lapied, Bernard Roy, Stéphane Deparis and two anonymous reviewers for very helpful comments.

References

M. Allais. The So-Called Allais Paradox and Rational Decisions under Uncertainty. In M. Allais and O. Hagen, editors, *Expected Utility Hypotheses and the Allais Paradox*, number 21 in Theory and Decision Library, pages 437–681. Springer, 1979. ISBN 978-94-015-7629-1. URL http://doi.org/10.1007/978-94-015-7629-1_17.

L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34 (1-3):197–215, 2002. ISSN 1012-2443. doi:10.1023/A:1014490210693.

L. Amgoud and H. Prade. Using arguments for making and explaining deci-

sions. *Artificial Intelligence*, 173(3–4):413–436, Mar. 2009. ISSN 0004-3702. doi:10.1016/j.artint.2008.11.006.

L. Amgoud, Y. Dimopoulos, and P. Moraitis. Making Decisions through Preference-Based Argumentation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference, KR 2008*, pages 113–123, Sydney, Australia, 2008. URL http://www.aaai.org/Library/KR/2008/kr08-012.php.

P. Baroni and M. Giacomin. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, 171(10–15):675–700, July 2007. ISSN 0004-3702. doi:10.1016/j.artint.2007.04.004.

P. Baroni and M. Giacomin. Semantics of Abstract Argument Systems. In G. Simari and I. Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 25–44. Springer, 2009. ISBN 978-0-387-98197-0. URL http://doi.org/10.1007/978-0-387-98197-0_2.

T. Bench-Capon and K. Atkinson. Abstract Argumentation and Values. In G. Simari and I. Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 45–64. Springer, Jan. 2009. ISBN 978-0-387-98196-3 978-0-387-98197-0. URL http://doi.org/10.1007/978-0-387-98197-0_3.

T. J. M. Bench-Capon. Persuasion in Practical Argument Using Value-based Argumentation Frameworks. *Journal of Logic and Computation*, 13(3):429–448, June 2003. ISSN 0955-792X. doi:10.1093/logcom/13.3.429.

P. Besnard and A. Hunter. *Elements of argumentation*. MIT Press, 2008. ISBN 978-0-262-02643-7. URL https://doi.org/10.7551/mitpress/9780262026437.001.0001.

G. Carenini and J. D. Moore. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170(11):925–952, Aug. 2006. ISSN 0004-3702. doi:10.1016/j.artint.2006.05.003.

F. Dietrich and C. List. A Reason-Based Theory of Rational Choice. *Noûs*, 47 (1):104–134, 2013. ISSN 1468-0068. doi:10.1111/j.1468-0068.2011.00840.x.

F. Dietrich and C. List. Reason-based choice and context-dependence: an explanatory framework. *Economics & Philosophy*, 32(2):175–229, 2016. ISSN 1474-0028. doi:10.1017/S0266267115000474.

Y. Dimopoulos, J.-G. Mailly, and P. Moraitis. Control Argumentation Frameworks. In S. A. McIlraith and K. Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16639.

P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321 – 357, 1995. ISSN 0004-3702. doi:10.1016/0004-3702(94)00041-X.

D. Ellsberg. Risk, Ambiguity, and the Savage Axiomes. *The Quarterly Journal of Economics*, 75:643–669, 1961. doi:10.2307/1884324.

D. Estlund. *Democratic Authority: A Philosophical Framework*. Princeton University Press, 2009. URL http://press.princeton.edu/titles/8571.html.

E. Ferretti, L. H. Tamargo, A. J. García, M. L. Errecalde, and G. R. Simari. An approach to decision making based on dynamic argumentation systems. *Artificial Intelligence*, 242:107–131, Jan. 2017. ISSN 0004-3702. doi:10.1016/j.artint.2016.10.004.

N. Goodman. *Fact, Fiction, and Forecast*. Harvard University Press, 4th edition, 1983. URL http://www.hup.harvard.edu/catalog.php?isbn=9780674290716.

J. Habermas. *Theorie des kommunikativen Handelns*. Suhrkamp, 1981. URL http://www.suhrkamp.de/buecher/theorie_des_kommunikativen_handelns-juergen_habermas_28775.html.

D. Hausman, M. *Preference, Value, Choice, and Welfare*. Cambridge University Press, 2011. URL https://doi.org/10.1017/CBO9781139058537.

A. Hunter. Probabilistic qualification of attack in abstract argumentation. *International Journal of Approximate Reasoning*, 55(2):607–638, 2014. ISSN 0888-613X. doi:10.1016/j.ijar.2013.09.002.

A. Kontoleon, U. Pascual, and T. M. Swanson, editors. *Biodiversity Economics*. Cambridge University Press, 2007. ISBN 978-0-521-86683-5. URL https://doi.org/10.1017/CBO9780511551079.

C. Labreuche. A general framework for explaining the results of a multi-attribute preference model. *Artificial Intelligence*, 175(7–8):1410 – 1448, 2011. ISSN 0004-3702. doi:10.1016/j.artint.2010.11.008.

M. Landry, J. Malouin, and M. Oral. Model validation in operations research. *European Journal of Operational Research*, 14:207–220, 1983. doi:10.1016/0377-2217(83)90257-6.

M. Landry, C. Banville, and M. Oral. Model legitimisation in operational research. *European Journal of Operational Research*, 92(3):443–457, 1996. doi:10.1016/0377-2217(96)00003-3.

K. R. MacCrimmon and S. Larsson. Utility Theory: Axioms Versus 'Paradoxes'. In M. Allais and O. Hagen, editors, *Expected Utility Hypotheses and the Allais Paradox*, number 21 in Theory and Decision Library, pages 333–409. Springer, 1979. ISBN 978-94-015-7629-1. URL http://doi.org/10.1007/978-94-015-7629-1_15.

M. J. Marcos, M. A. Falappa, and G. R. Simari. Dynamic Argumentation in Abstract Dialogue Frameworks. In P. McBurney, I. Rahwan, and S. Parsons, editors, *Argumentation in Multi-Agent Systems*, Lecture Notes in Computer Science, pages 228–247. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-21940-5. doi:10.1007/978-3-642-21940-5_14.

Y. Meinard, A. Remy, and B. Schmid. Measuring Impartial Preference for Biodiversity. *Ecological Economics*, 132:45–54, Feb. 2017. ISSN 0921-8009. doi:10.1016/j.ecolecon.2016.10.007.

O. Morgenstern. Some reflections on utility. In M. Allais and O. Hagen, editors, *Expected Utility Hypotheses and the Allais Paradox*, number 21 in Theory and Decision Library, pages 175–183. Springer, 1979. ISBN 978-94-015-7629-1. URL http://doi.org/10.1007/978-94-015-7629-1_6.

H. Prakken. Combining Sceptical Epistemic Reasoning with Credulous Practical Reasoning. In *Proceedings of the 2006 Conference on Computational Models of Argument: COMMA 2006*, pages 311–322, Amsterdam, The Netherlands, 2006. IOS Press. ISBN 978-1-58603-652-2. URL http://www.cs.uu.nl/groups/IS/archive/henry/b+a.pdf. Corrected version, April 2008.

H. Prakken. Models of Persuasion Dialogue. In G. Simari and I. Rahwan, editors, *Argumentation in Artificial Intelligence*, pages 281–300.

Springer, 2009. ISBN 978-0-387-98197-0. URL http://doi.org/10.1007/978-0-387-98197-0_14.

I. Rahwan and G. R. Simari, editors. *Argumentation in Artificial Intelligence*. Springer, 2009. URL http://doi.org/10.1007/978-0-387-98197-0.

H. Raiffa. Back from Prospect Theory to Utility Theory. In M. Grauer, M. Thompson, and A. P. Wierzbicki, editors, *Plural Rationality and Interactive Decision Processes*, number 248 in Lecture Notes in Economics and Mathematical Systems, pages 100–113. Springer, 1985. ISBN 978-3-540-15675-8. URL http://doi.org/10.1007/978-3-662-02432-4_8.

J. Rawls. *A Theory of Justice*. Harvard University Press, 1999. URL http://www.hup.harvard.edu/catalog.php?isbn=9780674000780.

J. Rawls. *Political Liberalism: Expanded Edition*. Columbia University Press, 2005. ISBN 978-0-231-52753-8. URL https://cup.columbia.edu/book/political-liberalism/9780231527538.

J. Rosenhead and J. Mingers, editors. *Rational Analysis for a Problematic World Revisited: Problem Structuring Methods for Complexity, Uncertainty and Conflict*. Wiley, 2nd edition, 2001. URL http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471495239.html.

N. D. Rotstein, M. O. Moguillansky, A. J. García, and G. R. Simari. A Dynamic Argumentation Framework. In P. Baroni, F. Cerutti, M. Giacomin, and G. R. Simari, editors, *Computational Models of Argument: Proceedings of COMMA 2010, Desenzano del Garda, Italy, September 8-10, 2010*, volume 216 of *Frontiers in Artificial Intelligence and Applications*, pages 427–438. IOS Press, 2010. ISBN 978-1-60750-618-8. doi:10.3233/978-1-60750-619-5-427.

B. Roy. *Multicriteria Methodology for Decision Aiding*. Kluwer Academic, 1996. ISBN 978-1-4757-2500-1. URL http://doi.org/10.1007/978-1-4757-2500-1.

L. J. Savage. *The Foundations of Statistics*. Courier Corporation, 2nd edition, 1972. ISBN 978-0-486-62349-8. URL http://store.doverpublications.com/0486623491.html.

T. M. Scanlon. *What We Owe to Each Other*. Harvard University Press, 2000. URL http://www.hup.harvard.edu/catalog.php?isbn=9780674004238.

A. Sen. *The idea of justice*. Harvard University Press, 2009. ISBN 978-0-674-06047-0. URL http://www.hup.harvard.edu/catalog.php?isbn=9780674060470.

P. Slovic and A. Tversky. Who accepts Savage's axiom? *Behavioral Science*, 19:368–373, 1974. doi:10.1002/bs.3830190603.

A. Proofs, and additional explanatory results

Our main goal in this section is to prove theorem 2. We do this by first proving that if a set S_γ is CAC, then it includes enough decisive arguments to settle the issue (we will call such a set $S \subseteq S^*$ *efficient*). This requires a few intermediate lemmas. Efficiency will bring a number of consequences of interest to us, among which theorem 2. As a second goal, we want to give some further results that help understand the relationship between the notions of clear-cut, validity and operational validity, existence of a CAC set of arguments, and efficiency.

Let us start with the formal definition of efficiency.

Definition 13 (Efficiency)**.** *Given a decision situation* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$ *and* $S \subseteq S^*$*, S is* efficient *iff* $T_i = \rightsquigarrow(S \cap \overline{\rhd_\exists(S^*)})$*, and* $t \notin T_i \Leftrightarrow \rightsquigarrow^{-1}(t) \subseteq \rhd_\forall(S \cap \overline{\rhd_\exists(S^*)})$.

Recall that $\overline{\rhd_\exists(S^*)}$ designates the arguments not trumped by any argument, thus, the decisive arguments, and hence, $\rhd_\forall(S \cap \overline{\rhd_\exists(S^*)})$ designates the arguments always trumped by some decisive argument in S.

In all this section, we assume we are given a decision situation $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$ and a subset of arguments $S_\gamma \subseteq S^*$ (except in theorem 5).

Our strategy for proving that CAC implies efficiency, roughly speaking, involves excluding "undecided" situations from S_γ. For example, we want to show that it is impossible that an argument has no decisive argument trumping it in S_γ, but also fails to be defended in S_γ. We will do this by progressively promoting or degrading arguments, e.g., show that, in S_γ, if an argument is resistant (has no argument that decisively trumps it), then it must also be defended, and if it is defended, it must be replaceable by decisive arguments.

Define $S_{\gamma\text{dec}} = S_\gamma \cap \overline{\rhd_\exists(S^*)}$ as the decisive arguments from S_γ.

Define an argument s as finitely defended iff some finite set of arguments from $S_{\gamma\text{dec}}$ defends it, thus, iff $\exists S \subseteq S_{\gamma\text{dec}}$ such that $\rhd_\exists^{-1}(s) \subseteq \rhd_\exists(S)$, S finite. Define $S_{\gamma\text{def}}$ as the arguments from S_γ that are finitely defended.

Define $R_{\gamma\text{dec}} \subseteq S^*$ as the arguments that are replaceable by $S_{\gamma\text{dec}}$. Recall that S replaces s iff $\rhd_\exists(s) \subseteq \rhd_\exists(S)$ and $\rightsquigarrow(s) \subseteq \rightsquigarrow(S)$.

Define $S_{\gamma\text{res}} = S_\gamma \cap \overline{\rhd_\exists(S_{\gamma\text{dec}})}$ as the resistant arguments from S_γ, namely, those not trumped by any argument from $S_{\gamma\text{dec}}$.

Define $E_{\gamma\text{res}} \subseteq S^*$ as the arguments that are essentially replaceable by $S_{\gamma\text{res}}$. Recall that S essentially replaces s iff $(\rhd_\exists(s) \cap S_\gamma) \subseteq \rhd_\exists(S)$ and $\rightsquigarrow(s) \subseteq \rightsquigarrow(S)$.

Similarly, $E_{\gamma\text{dec}}$ are the arguments essentially replaceable by $S_{\gamma\text{dec}}$.

Lemma 1 ($S_{\gamma\text{def}} \subseteq R_{\gamma\text{dec}}$)**.** *If S_γ is Closed under reinstatement and Answerable, then the arguments from S_γ that are finitely defended are replaceable by decisive arguments from S_γ; formally: $S_{\gamma\text{def}} \subseteq R_{\gamma\text{dec}}$.*

Proof. The strategy for this proof is the following. If $s \in S_{\gamma\text{def}}$, some finite set of arguments defends s. We wish to pick defenders one by one, replacing s by applying Closed under reinstatement to s and the chosen defender, obtaining an argument that fewer arguments trump, and then show that iterating the process yields a decisive argument replacing s.

We need the following intermediate result. Assume a set of arguments $S \subseteq S_{\gamma\text{dec}}$ is given, together with an argument $s_1 \in S$ and an argument $s_1^{\text{r}} \in S_\gamma$ defended by S. Then, there exists an argument $s_2^{\text{r}} \in S_\gamma$ replacing s_1^{r} and defended by $S \setminus \{s_1\}$.

Indeed, from Answerability, because $s_1 \in S_{\gamma\text{dec}}$, $\rhd_\exists(s_1) = \rhd_\forall(s_1)$. Also, as $s_1 \in S_{\gamma\text{dec}}$, we can assume that $s_1 \neq s_1^{\text{r}}$, otherwise $s_1^{\text{r}} \in S_{\gamma\text{dec}}$ and the result is obtained by taking $s_2^{\text{r}} = s_1^{\text{r}}$. And s_1 does not trump s_1^{r}, otherwise s_1^{r} is trumped by a decisive argument and thus not defended. We can thus apply Closed under reinstatement to (s_1, s_1^{r}). We obtain that for some $s_2^{\text{r}} \in S_\gamma$, s_2^{r} replaces s_1^{r} and $\rhd_\exists^{-1}(s_2^{\text{r}}) \subseteq \rhd_\exists^{-1}(s_1^{\text{r}}) \setminus \rhd_\exists(s_1)$. Thus, $S \setminus \{s_1\}$ defends s_2^{r}: any argument trumping s_2^{r} already trumped s_1^{r}, hence, is trumped by S (because that set defends s_1^{r}), and is not trumped by s_1. This proves our intermediate result.

Coming back to the main point, we know that a finite coalition $S \subseteq S_{\gamma\text{dec}}$ defends $s \in S_\gamma$. Define $s_1^{\text{r}} = s$ and apply the intermediate result repetitively to obtain an argument $s_2^{\text{r}} \in S_\gamma$ replacing s and defended by S minus one element, then $s_3^{\text{r}} \in S_\gamma$ replacing s_2^{r}, thus, replacing s (because replacement is transitive) and defended by S minus two elements, and so on, until obtaining a replacer defended by $\emptyset$, thus, decisive. □

Lemma 2 ($S^* = E_{\gamma\text{res}} \cup \rhd_\exists(S_{\gamma\text{res}})$)**.** *If S_γ is covering, any argument is either essentially replaceable by $S_{\gamma\text{res}}$, or attacked by an argument from $S_{\gamma\text{res}}$; formally: $S^* = E_{\gamma\text{res}} \cup \rhd_\exists(S_{\gamma\text{res}})$.*

Proof. We consider in turn three sets whose union yields S^*: $\overline{S_\gamma}$, $S_\gamma \cap \rhd_\exists(S_{\gamma\text{dec}})$ and $S_\gamma \cap \overline{\rhd_\exists(S_{\gamma\text{dec}})}$.

First, $\overline{S_\gamma} \subseteq E_{\gamma\text{res}} \cup \rhd_\exists(S_{\gamma\text{res}})$: from covering, if $s \notin S_\gamma$, s is unnecessary, and by definition, s is unnecessary iff $s \in E_{\gamma\text{res}}$ or $s \in \rhd_\exists(S_{\gamma\text{res}})$.

Second, $S_\gamma \cap \rhd_\exists(S_{\gamma\text{dec}}) \subseteq \rhd_\exists(S_{\gamma\text{dec}}) \subseteq \rhd_\exists(S_{\gamma\text{res}})$, because $S_{\gamma\text{dec}} \subseteq S_{\gamma\text{res}}$.

Third, $S_\gamma \cap \overline{\rhd_\exists(S_{\gamma\text{dec}})} \subseteq E_{\gamma\text{res}}$, because $S_\gamma \cap \overline{\rhd_\exists(S_{\gamma\text{dec}})} = S_{\gamma\text{res}}$ by definition.

We have considered all three possible cases, and the conclusion obtains in all cases. □

Lemma 3 ($S_{\gamma\text{res}} \subseteq S_{\gamma\text{def}}$)**.** *If S_γ is CAC, any argument in S_γ that has no argument that decisively trumps it is finitely defended; formally: $S_{\gamma\text{res}} \subseteq S_{\gamma\text{def}}$.*

Proof. Recall that the relation Q is defined in Bounded length (condition 9) as $Q = (\rhd_\exists \cup (\rhd_\exists \circ \rhd_\exists)) \cap (S_\gamma \times S_\gamma)$. Observe that, given any set $S \neq \emptyset$, Bounded Length forbids that $\forall s \in S : S \cap Q^{-1}(s) \neq \emptyset$. Otherwise, applying Q^{-1} to an element of S would always yield some element in S, and Q^{-1} could then be applied any desired number of times starting from any $s \in S$, thereby building a chain as long as desired. Accordingly, for any set S, Bounded Length imposes that if $\forall s \in S : S \cap Q^{-1}(s) \neq \emptyset$, then $S = \emptyset$.

Define $S = S_{\gamma\text{res}} \cap \overline{S_{\gamma\text{def}}}$. We show that, given any $s \in S$, $S \cap Q^{-1}(s) \neq \emptyset$. This suffices to obtain $S = \emptyset$ and, therefore, our desired conclusion.

Pick any $s \in S$. Towards exhibiting an argument in $S \cap Q^{-1}(s)$, we want first to exhibit some argument s' that is a) trumped by some argument $s^* \in S_{\gamma\text{res}}$, thus $s' \in \rhd_\exists(S_{\gamma\text{res}})$; b) not trumped by any argument in $S_{\gamma\text{dec}}$, thus $s' \notin \rhd_\exists(S_{\gamma\text{dec}})$; c) equal to s or trumping s. As a second step, from the existence of such an s' we will then prove that s^*, the particular trumping argument in part a), belongs to S (thanks to parts a) and b)), and belongs to $Q^{-1}(s)$ (thanks to part c)).

Our first step thus amounts to show that some s' satisfies our three conditions above.

From $s \notin S_{\gamma\text{def}}$ and $s \in S_\gamma$, we know that s is not finitely defended, and using the contrapositive of Bounded width, we obtain that s is not infinitely defended either. Hence, by definition of defense, there exists some $s_1 \in \overline{\rhd_\exists(S_{\gamma\text{dec}})} \cap \rhd_\exists^{-1}(s)$. And, applying $[S^* = E_{\gamma\text{res}} \cup \rhd_\exists(S_{\gamma\text{res}})]$, either $s_1 \in E_{\gamma\text{res}}$, or $s_1 \in \rhd_\exists(S_{\gamma\text{res}})$.

If $s_1 \in E_{\gamma\text{res}}$, $s \in \rhd_\exists(S_{\gamma\text{res}})$. Besides, because $s \in S$, $s \in S_{\gamma\text{res}}$. Thus taking $s' = s$ satisfies our three conditions.

And if $s_1 \in \rhd_\exists(S_{\gamma\text{res}})$, because $s_1 \in \overline{\rhd_\exists(S_{\gamma\text{dec}})}$), taking $s' = s_1$ satisfies our three conditions.

For our second step, consider an argument $s^* \in S_{\gamma\text{res}}$ that trumps s' (we know this is possible thanks to part a)). Thanks to part b), we know that s' is not trumped by any argument in $S_{\gamma\text{dec}}$, and from $[S_{\gamma\text{def}} \subseteq R_{\gamma\text{dec}}]$, we know

that if s' was trumped by an argument in $S_{\gamma\text{dec}}$, it would be trumped by an argument in $S_{\gamma\text{dec}}$, thus, s' is not trumped by any argument in $S_{\gamma\text{def}}$. Because $s^* \rhd_\exists s'$, we know that $s^* \notin S_{\gamma\text{def}}$. Thus, $s^* \in S$. Finally, $s^* \rhd_\exists s$ or $s^* \rhd_\exists s' \rhd_\exists s$ (thanks to part c)), thus, $s^* \in Q^{-1}(s)$. □

Lemma 4 ($S^* = E_{\gamma\text{dec}} \cup \rhd_\exists(S_{\gamma\text{dec}})$)**.** *If S_γ is CAC, any argument is either essentially replaceable by decisive arguments from S_γ, or attacked by a decisive argument from S_γ; formally: $S^* = E_{\gamma\text{dec}} \cup \rhd_\exists(S_{\gamma\text{dec}})$.*

Proof. This follows from $[S^* = E_{\gamma\text{res}} \cup \rhd_\exists(S_{\gamma\text{res}})]$, $[S_{\gamma\text{res}} \subseteq S_{\gamma\text{def}}]$ and $[S_{\gamma\text{def}} \subseteq R_{\gamma\text{dec}}]$. □

Theorem 3 (CAC implies efficiency)**.** *If S_γ is CAC, S_γ is efficient.*

Proof. We prove that $\rightsquigarrow(E_{\gamma\text{dec}}) \subseteq \rightsquigarrow(S_{\gamma\text{dec}}) \subseteq T_i \subseteq \rightsquigarrow(\overline{\rhd_\forall(S_{\gamma\text{dec}})}) \subseteq \rightsquigarrow(E_{\gamma\text{dec}})$.
This proves the point, as it shows that

i. $T_i = \rightsquigarrow(S_{\gamma\text{dec}})$, and

ii. $t \notin T_i \Leftrightarrow \rightsquigarrow^{-1}(t) \subseteq \rhd_\forall(S_{\gamma\text{dec}})$, because $T_i = \rightsquigarrow(\overline{\rhd_\forall(S_{\gamma\text{dec}})})$.

That $\rightsquigarrow(E_{\gamma\text{dec}}) \subseteq \rightsquigarrow(S_{\gamma\text{dec}}) \subseteq T_i$ follows from the definitions of $E_{\gamma\text{dec}}$ and T_i.
The next subset relation holds because if some decisive argument supports t, that argument is not in $\rhd_\forall(S_{\gamma\text{dec}})$.
Finally, Answerability mandates that $\rhd_\exists(S_{\gamma\text{dec}}) \subseteq \rhd_\forall(S_{\gamma\text{dec}})$, from which it follows that $\overline{\rhd_\forall(S_{\gamma\text{dec}})} \subseteq \overline{\rhd_\exists(S_{\gamma\text{dec}})}$, and using $[S^* = E_{\gamma\text{dec}} \cup \rhd_\exists(S_{\gamma\text{dec}})]$, $\overline{\rhd_\exists(S_{\gamma\text{dec}})} \subseteq E_{\gamma\text{dec}}$. □

Theorem 4 (Validity of η)**.** *Assume S_γ is efficient and η, a model of the decision situation, is S_γ-operationally valid. Then $T_i = T_\eta$.*

Proof. Recall that a model is *S_γ-operationally valid* iff for all $(s \rightsquigarrow_\eta t)$, $s \in S^*$, we have $[s \rightsquigarrow t]$ and $[\forall s_c \in S_\gamma : (s_c \not\rhd_\exists s) \vee (\exists s_{cc} \in S^* \mid s_{cc} \rhd_\eta s_c \wedge s_{cc} \rhd_\exists s_c)]$, and when t is not supported by η, $\forall s \in S_\gamma \mid s \rightsquigarrow t : (\exists s_c \in S^* \mid s_c \rhd_\eta s \wedge s_c \rhd_\exists s)$.
Consider $t \in T_\eta$. By definition, some $s \rightsquigarrow_\eta t$. From operational validity of η, we obtain that $s \rightsquigarrow t$ and $\forall s_c \rhd_\forall s : s_c \notin S_\gamma \cap \overline{\rhd_\exists^{-1}(S^*)}$ (because $[s_c \rhd_\forall s \wedge s_c \in S_\gamma] \Rightarrow s_c \in \rhd_\exists^{-1}(S^*)$). Hence, $s \notin \rhd_\forall(S_\gamma \cap \overline{\rhd_\exists^{-1}(S^*)})$, thus $\rightsquigarrow^{-1}(t) \not\subseteq \rhd_\forall(S_\gamma \cap \overline{\rhd_\exists^{-1}(S^*)})$. Efficiency of S_γ brings $t \in T_i$.
If $t \notin T_\eta$, from operational validity of η, no decisive argument in S_γ may support t, equivalently, $t \notin \rightsquigarrow(S_\gamma \cap \overline{\rhd_\exists(S^*)})$, and from efficiency, $t \notin T_i$. □

We can now prove theorem 2.

Proof of theorem 2. From [CAC implies efficiency], we obtain that S_γ is efficient. It then follows from the efficiency of S_γ that the decision situation is clear-cut and that a S_γ-operationally valid model exists. The last consequence is given by theorem 4. □

The following theorem may help clarify the relationship between efficiency, existence of CAC arguments, and the situation admitting a model as we conceive it.

Theorem 5 (CAC subset equivalent to efficiency)**.** *Given a decision situation* $(T, S^*, \rightsquigarrow, \rhd_\exists, \not\rhd_\exists)$ *and a subset of arguments* $S \subseteq S^*$*, there exists a set* $S_\gamma \subseteq S$ *that is CAC iff* S *is efficient.*

Proof. From [CAC implies efficiency], if some set $S_\gamma \subseteq S$ is CAC, then S_γ is efficient, and because efficiency propagates to supersets, S is efficient.

If S is efficient (thus, the decision situation is clear-cut), then a CAC subset S_γ exists: suffices to choose as members of S_γ only the decisive arguments required to support the justifiable propositions and trump the supporters $s \rightsquigarrow t$ of untenable propositions. Observing that no arguments trump any argument in the resulting set (thus $s \rhd_\exists s_\gamma$ for no $s \in S^*, s_\gamma \in S_\gamma$), most of the conditions for S_γ to be CAC are immediately seen to be satisfied. About arguments $s \in S^* \setminus S_\gamma$ being unnecessary, we only have to show that when $s \rightsquigarrow t$, either s is trumped by an argument from S_γ that is not decisively trumped, or s is essentially replaceable by arguments from S_γ. Indeed, by our construction of S_γ, if s supports an accepted t, it is essentially replaceable, and otherwise, it is trumped by a decisive argument. □