Explainable Decisions under Incomplete Knowledge with Supports and Weights

Florence Dupin de Saint-Cyr, Romain Guillaume, Umer Mushtaq

Explainable Decisions under Incomplete Knowledge with Supports and Weights*

Florence Dupin de Saint-Cyr
IRIT, Toulouse University
Toulouse, France
bannay@irit.fr

Romain Guillaume
IRIT, Toulouse University
Toulouse, France
romain.guillaume@irit.fr

Umer Mushtaq
LABEX-MME DII / LEMMA
Univ. Paris II - Pantheon Assas,Paris, France
umer.mushtaq@u-paris2.fr

May 30, 2019

Abstract

Our research concerns the problem of explainable decision in a context of incomplete knowledge. We define a framework called Bipolar Layered Framework with Support and Weights (BLFSW) that represents the set of argument graphs that can be used in the domain, enabling us to compute what results can be obtained in the different decision situations. This framework also contains information about the utilities/disutilities of these tangible results. This paper extends Bipolar Layered Frameworks defined in [1] by enabling the expression of *supports* for decision principles and by giving the user the possibility to fix the strength of inhibitors and supports with *weights*. This increased expressiveness of the framework is important both for refining the evaluation of alternatives and to improve the compactness of the representation. The main result of this paper is to provide an automatic way to explain a possibilistic decision setting in terms of a BLFSW which makes explicit the principles that govern the decision.

Keywords: Possibility theory, qualitative decision theory, explanation.

1 Introduction

An individual decision analysis process involves an agent (the decision maker) taking account of the decision situation and then evaluating different courses of action (decision alternatives). In order to be able to do so, the decision maker must characterize the decision-making situation with respect to two distinct components: a formulation of the decision goals and a characterization of the decision alternatives [2]. Usually, the evaluation is based on an associated utility function (see e.g.

*This is a draft version, the paper is published in the proceedings of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2019), New Orleans, Louisiana, USA, 23/06/19-26/06/19, Alina Zare, Derek Anderson (Eds.), June 2019.

the introductory book of [3]) which encodes the satisfaction degrees reached by choosing each decision alternative. Despite a lot of works on decision theory, two issues are often not easy to solve. The information of the agent about the decision situation is often uncertain, incomplete and distributed. Hence the first issue is to deal with imperfect information (uncertainty, incomplete and distributed knowledge). The second issue is to be able to explain and justify the decisions that are made. It is also a desirable goal to enable the decision makers to have a broader view of the principles that govern the decision and to enable them to participate in their elaboration. These issues are even more important when the decision to be made concerns a group of autonomous agents that have their own knowledge and preferences.

Classical qualitative decision-making approaches use aggregation criteria that combine the measurement of uncertainty with utility (see e.g. [4]). Roughly speaking we can sum up the standard approaches of decision under uncertainty as follows: the decision maker defines a utility function $f(d, s)$ which associates a value to a decision d in a given scenario s. The second step consists in defining an aggregation function on all the possible scenarios given the uncertain current knowledge about the real situation. We propose an approach which allows the user to choose the best decision and also to explain it. Indeed, instead of using a utility function over all possible states, we propose to use a set of decision principles (DP). A DP relates some characteristic features of the situation to the achievement of a tangible result (which has a utility level). For instance, if an agent wants to find a hotel, we can enunciate a decision principle saying that "a priori, a hotel with a pool gives the opportunity to swim" (where "swimming" is a tangible result with a good level of utility for our agent). Our approach is a two step process: the first step computes the certainty of the achievement of some tangible results, leading to compute the function $N(r|S)$ that gives the necessity of having the result r given a set of situations S (obtained by a decision made on an uncertain scenario). Then the second step consists in aggregating (taking into account their importance, polarities and certainty) the possible results that can be obtained in a given situation in order to compare the different situations. For this step, we use an extension of the qualitative bipolar approach of Dubois and Fargier [5] where positive arguments (pros) and negative arguments (cons) are uncertain.

The particular originality of the BLFSW is mainly its first step process which is done by using several argumentation graphs, each argumentation graph enables us to assess about the achievement of one tangible result. In argumentation theory, there are two kinds of actions on arguments: *attacks* that tend to say that the conclusion of the argument (here the tangible result) is not achieved in a given situation and *supports* that increase the belief degree that the result is achieved. Principles in BLFSWs are akin to arguments in that they state a reason for believing that a tangible result is obtained. The notion of *argument* in favor or against a decision has been developed in practical argumentation domain which has been widely studied (see e.g. [6, 7]) since the initial proposal of Raz [8] and the philosophical justification provided by Walton [9]. Practical argumentation aims at answering

the question *what is the right thing to do in a given situation* which is clearly related to a decision problem. Several works are using argumentative approaches to tackle it: Amgoud and Prade [10] propose a bipolar argumentation-based approach distinguishing epistemic and practical arguments. Argumentation has also been proposed to govern decision making in a negotiation context (see for instance [11] and [12] for a survey). However in all the argumentative approaches mentioned above, it is difficult to obtain a precise explanation of the decision: either because the arguments are abstract and only the attack relation between them is informative, or because there is no clear explanation of how and why the content of the argument justifies the final choices. BLFSW is a new method for reasoning about decision arguments in which more place is given to explanation by making explicit both the decision principles and their supports and inhibitors.

2 BLF with supports and weights

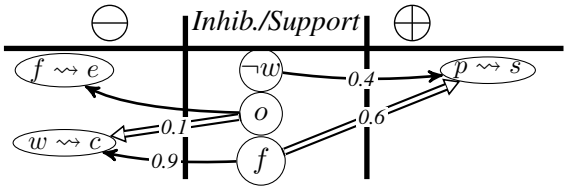
A BLFSW is a visual bipolar framework that represents all explicit information known about a decision domain. Hence, it contains both the knowledge for reasoning about the achievement of tangible results (called goals) and the preference information associated to these goals: namely their polarities and their importance level. The polarity of a goal is positive if it is a desirable result, it is negative when this result should be avoided.

We consider a set $\mathscr{A}$ of alternatives about which some information is available and two languages $\mathscr{L}_F$ (a propositional language based on a vocabulary $\mathcal{V}_F$) representing information about some features that are believed to hold for an alternative and $\mathscr{L}_G$ (another propositional language based on a distinct vocabulary $\mathcal{V}_G$) representing information about the achievement of some goals when an alternative is selected. In the propositional languages used here, the logical connectors *or*, *and*, *not* are denoted respectively by $\vee$, $\wedge$, and $\neg$. A *literal* is a propositional symbol x or its negation $\neg x$, the set of literals of $\mathscr{L}_G$ are denoted by LIT_G. Classical inference, logical equivalence and contradiction are denoted respectively by $\models$, $\equiv$, $\bot$. We use a special symbol $\rightsquigarrow$ to encode an a priori deduction, called Decision Principle, from some observations to a goal.

Example 1. *Let us imagine an agent who wants to find a hotel which is not expensive (e) and in which he can swim (s). This agent prefers to avoid crowded hotels (c). The possible pieces of information concern the following attributes: $\mathcal{V}_F = \{p, f, w, o\}$ that describes the respective features of the hotel "to have a pool", "to be a four star hotel", "to be in a place where the weather is fine", "to propose special offers". The agent may consider the following principles: $\mathcal{P} = \{p \rightsquigarrow s, f \rightsquigarrow e, w \rightsquigarrow c\}$. They respectively express that "a priori when there is a pool the agent can swim", "a priori if the hotel is four star then it is expensive" and "if the weather is fine in this area then the hotel is a priori crowded".*

The principles can be supported or inhibited, this is represented by double and

single arcs that are weighted accordingly to the strength of the support/inhibition[1]*. For instance, a special offer increases the certainty to have a crowded hotel when the weather is fine. Four star hotels are expensive but a special offer may inhibit this deduction. The following picture is a graphical representation of this example by a BLFSW: it is a tripartite graph represented in three columns, the DPs with a positive goal are situated on the left column, the inhibitors and supports are in the middle, and the DPs with a negative polarity are situated on the right. The more important (positive and negative) DPs are in the higher part of the graph, equally important DPs are drawn at the same horizontal level. Hence the highest positive level is at the top left of the figure, the bottom right contains DPs with negative goals of low importance. The heights of the inhibitors and supports are not significant; only their existence is meaningful.*



More formally, a BLFSW is defined as follows:

Definition 1 (BLFSW)**.** *A Bipolar Layered Framework with Supports and Weights is a tuple* $(\mathcal{P}, \mathcal{I}, \mathcal{S}, pol, \preceq, w)$*.* $\mathcal{P}$ *is a set of decision principles:* $\mathcal{P} = \{\varphi \rightsquigarrow g | \varphi \in \mathscr{L}_F, g \in LIT_G\}$*.* $\mathcal{I} \subseteq (\mathcal{L}_F \times \mathcal{P})$ *is a set of inhibitors.* $\mathcal{S} \subseteq (\mathcal{L}_F \times \mathcal{P})$ *is a set of supports.* pol *is a function* $pol : \mathcal{V}_G \rightarrow \{\oplus, \ominus\}$ *which gives the polarity of a goal* $g \in \mathcal{V}_G$*, this function is extended to goal literals by* $pol(\neg g) = -pol(g)$ *with* $-\oplus = \ominus$ *and* $-\ominus = \oplus$ *and to DPs accordingly:* $pol(\varphi, g) = pol(g)$*.* LIT_G *is totally ordered by the relation* $\preceq$ *("less or equally important than") and DPs are ordered accordingly:* $(\varphi \rightsquigarrow g) \preceq (\psi \rightsquigarrow g')$ *iff* $g \preceq g'$*.* $w : \mathcal{I} \cup \mathcal{S} \rightarrow]0, 1]$ *is a weight function on inhibitors and supports.*

The weight on a support/inhibitor of a DP is expressing an increased/decreased certainty degree about the fact that triggering this DP will lead to the achievement of its conclusion. We do not allow for supports or inhibitors of weight 0, since it would mean that there is no information about the supporting/inhibiting effects.

2.1 Reasoning about goal achievements

The first part of the process is a reasoning part: it consists in reasoning with the argumentation graphs that concern each goal in order to check what are the realized goals. This is done by considering what is known: given a consistent knowledge base K, we first define a K-BLFSW as the BLFSW that is obtained when all what is known is K. More formally,

[1]For a simpler representation, the drawing of a BLFSW obeys the convention that *if no weight is given for a set of supports and inhibitors concerning the same DP then all weights are equal to* 1.

Definition 2 (K-BLFSW)**.** *Given a consistent knowledge base K and a BLFSW $B = (\mathcal{P}, \mathcal{I}, \mathcal{S}, pol, \preceq, w)$, a K-BLFSW associated to B is a tuple $(\mathcal{P}_K, \mathcal{I}_K, \mathcal{S}_K, pol, \preceq, w_K)$ where*

- $\mathcal{P}_K = \{(\varphi, g) \in \mathcal{P}, \text{ s.t. } K \models \varphi\}$ *is the set of valid DPs in $\mathcal{P}$ (i.e., those whose reason φ holds in K).*
- $\mathcal{I}_K = \{(\varphi, p) \in \mathcal{I}, \text{ s.t. } K \models \varphi \text{ and } p \in \mathcal{P}_K\}$ *is the set of* valid inhibitions *according to K.*
- $\mathcal{S}_K = \{(\varphi, p) \in \mathcal{S}, \text{ s.t. } K \models \varphi \text{ and } p \in \mathcal{P}_K\}$ *is the set of* valid supports *according to K.*
- w_K *is the restriction of w on $\mathcal{I}_K \cup \mathcal{S}_K$.*

The DPs that are not inhibited in the K-BLFSW are the ones that are trusted, in order to know if a DP is inhibited, we have to compare the weights of its inhibitors and supports.

Definition 3. *Given a K-BLFSW $(\mathcal{P}_K, \mathcal{I}_K, \mathcal{S}_K, pol, \preceq, w_K)$, we define the activation level of $p \in \mathcal{P}_K$ as follows:*

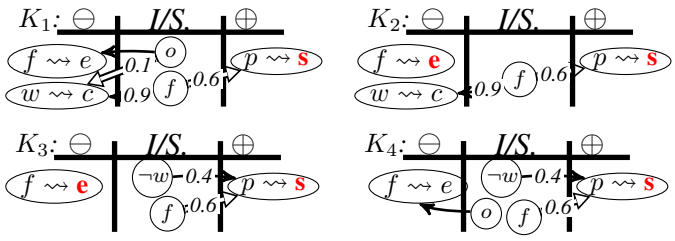
$$\alpha(p) = \sum_{s \in \mathcal{S}_K(p)} w_K(s, p) - \sum_{i \in \mathcal{I}_K(p)} w_K(i, p)$$

with $\mathcal{S}_K(p) = \{\psi \in \mathscr{L}_F | (\psi, p) \in \mathcal{S}_K\}$, $\mathcal{I}_K(p) = \{\psi \in \mathscr{L}_F | (\psi, p) \in \mathcal{I}_K\}$. According to $\alpha(p)$, the DP p is either inhibited *iff $\alpha(p) < 0$,* supported *iff $\alpha(p) > 0$ or* unaffected *iff $\alpha(p) = 0$.*

In other words, the weights of supports and inhibitors that concerns a given DP p are used to determine whether p is globally supported or inhibited or unaffected. The DP is said supported when supports are stronger than inhibitors, it is inhibited in the opposite case. When inhibitors and supports are equal they cancel each other.

Definition 4 (realized goals)**.** *Given a K-BLFSW $B_K = (\mathcal{P}_K, \mathcal{I}_K, \mathcal{S}_K, pol, \preceq, w_K)$, a goal g in LIT_G is said to be* realized *wrt B_K if there is a DP in $\mathcal{P}_K$ that concludes g and that is not inhibited.*

Example 1 (cont.): *Let us consider $K_1 = \{p \wedge w \wedge f \wedge o\}$, $K_2 = \{p \wedge w \wedge f\}$, $K_3 = \{p \wedge \neg w \wedge f\}$, $K_4 = \{p \wedge \neg w \wedge f \wedge o\}$. The four corresponding BLFSWs are:*



In K_1, $f \rightsquigarrow e$ is inhibited (since it has only one inhibitor with a default weight of 1), $w \rightsquigarrow c$ is inhibited (since it has an inhibitor of weight 0.9 which is heavier than the weight 0.1 of its support) and $p \rightsquigarrow s$ is supported. Hence the only realized goals of K_1 is s. Similarly, we compute the realized goals of the other hotels: K_2 and K_3 have the same realized goals: e and s, K_4 has only one realized goal: s (since the $p \rightsquigarrow s$ is supported by a support that is heavier than its inhibitor).

2.2 Handling preferences

Once the reasoning step is done, we know what goals are realized in what alternative situation, then the second step consists in taking into account the preferences expressed in terms of importance and polarities of the goals. Hence, the definition of realized goals wrt a K-BLFSW allows us to compare alternatives according to the goals they achieve.

In [13], three decision rules called Pareto, Bipolar Possibility and Bipolar Leximin have been introduced. We have chosen to only translate the Bipolar Leximin rule in order to compare two alternatives.

Definition 5 (BiLexi decision rule)**.** *Given a BLFSW $B = (\mathcal{P}, \mathcal{I}, \mathcal{S}, pol, \preceq, w)$ and two alternatives described respectively by K and K' with their associated realized goals $\mathtt{R}$ and $\mathtt{R}'$, the Bipolar Leximin dominance relation $\succeq_{BiLexi}$ ($BiLexi$-preferred to) is s.t.: let $X_g^{pol} = \{g' \in X \text{ s.t. } g \simeq g' \text{ and } pol(g') = pol\}$ and $M = \max(\{g \in \mathtt{R} \cup \mathtt{R}' \text{ s.t. } |\mathtt{R}_g^{\oplus}| \neq |\mathtt{R}_g'^{\oplus}| \text{ or } |\mathtt{R}_g^{\ominus}| \neq |\mathtt{R}_g'^{\ominus}|\}, \preceq)$*

$$\begin{array}{lll} K \succ_{BiLexi} K' & \textit{iff} & \left| \begin{array}{l} M \textit{ exists and} \\ |\mathtt{R}_M^{\oplus}| \geq |\mathtt{R}_M'^{\oplus}| \textit{ and } |\mathtt{R}_M^{\ominus}| \leq |\mathtt{R}_M'^{\ominus}| \end{array} \right. \\ K \simeq_{BiLexi} K' & \textit{iff} & M \textit{ does not exist} \end{array}$$

In other words, an alternative described by K is $BiLexi$-preferred to another one described by K' if there is a goal M such that the number of realized positive and negative goals at levels strictly more important than M are the same for K', but at the level M either the number of positive goals of K is greater than those of K' or the number of negative goals of K is lower than those of K'.

Example 1 (cont.): *The ranks of the hotels are:*

$$(K_1 \simeq_{BiLexi} K_4) \succ_{BiLexi} (K_2 \simeq_{BiLexi} K_3)$$

Note that in case of equality between two alternatives, the activation levels of the DPs that are justifying the goals achieved by the alternatives, can be used to choose between them.This means that goals are associated with two evaluations, one concerning their importance (that can be called utility when the goal is positive and disutility when it is negative, in our framework it is characterized by $\preceq$ and pol) and one concerning the certainty α_K about their realization in K, as defined below.

Definition 6 (Certainty of a realized goal)**.** *Given a BLFSW* $B = (\mathcal{P}, \mathcal{I}, \mathcal{S}, pol, \preceq, w)$ *and an alternative described by* K*, for all goal* g *realized in* B_K*, the certainty associated to* g *is:*

$$\alpha_K(g) = \max_{\varphi \in \mathscr{L}_F, p = \varphi \rightsquigarrow g \in \mathcal{P}_K} \alpha(p)$$

In other words, the certainty associated to g corresponds to the maximum activation level of a DP concluding g.

Definition 7 (BW decision rule)**.** *Given a BLFSW* $B = (\mathcal{P}, \mathcal{I}, \mathcal{S}, pol, \preceq, w)$ *and two alternatives described respectively by* K *and* K' *with their associated realized goals* $\mathtt{R}$ *and* $\mathtt{R}'$*, the BW dominance relation* $\succeq_{BW}$ *(*$BiLexi\&Weight$-preferred to*) is defined by:* $K \succ_{BW} K'$ *iff* $K \succ_{BiLexi} K'$ *or (*$K \simeq_{BiLexi} K'$ *and*
$\exists M = \max(\{g \in \mathtt{R} \cup \mathtt{R}' \mid \begin{matrix} \max_{g_1 \in \mathtt{R}_g^{\oplus}} \alpha_K(g_1) > \max_{g_2 \in \mathtt{R}'^{\oplus}_g} \alpha_{K'}(g_2) \text{ or} \\ \max_{g_1 \in \mathtt{R}_g^{\ominus}} \alpha_K(g_1) < \max_{g_2 \in \mathtt{R}'^{\ominus}_g} \alpha_{K'}(g_2) \end{matrix} \}, \preceq)$ *)*

In the previous definition, M is the highest important goal s.t. the maximum weight of a positive or a negative achieved goal of same priority for K and K' differs in favor of K i.e., either the maximum weight of positive achieved goals for K is strictly greater than the one for K' or the maximum weight of negative achieved goals for K is strictly lower than the one for K'.

Example 1 (cont.): *We get:* $K_1 \succ_{BW} K_4 \succ_{BW} K_2 \succ_{BW} K_3$*. Since in* K_1*,* $p \rightsquigarrow s$ *is supported and not attacked hence the activation level of* $p \rightsquigarrow s$ *is* 0.6*, while in* K_4*,* $p \rightsquigarrow s$ *has an activation level of* $0.6 - 0.4 = 0.2$ *which means that the achievement of swim is more certain in the situation described by* K_1 *than in the situation described by* K_4*. The same refinement is done to differentiate* K_2 *and* K_3*, their negative goal* e *is achieved with the same certainty while the positive goal* s *is more certainly achieved in* K_2 *than in* K_3*.*

3 Towards an automatic Explanation of a Possibilistic decision setting

As seen above, a BLFSW is a tool that enables the user to make explicit the decision setting. This paper aims at translating classical decision settings into BLFSW in order to give an automatic explanation to the utilities attached to decisions. In this section, we first write a reminder about Possibility Theory and Defaults, then we show how the decision principles, inhibitors and weights of a BLFSW can be interpreted in terms of possibility theory. This is done by following up the work of [14] in order to build DPs from uncertain knowledge expressed under the form of a possibility distribution on worlds and from preferences expressed as utilities associated to goals. In this process, a DP $\varphi \rightsquigarrow g$ is viewed as a defeasible rule saying that *if* φ *holds then a priori* g *is achieved*, and we explain how weighted inhibitions and supports can be defined according to this view. The third subsection show how to automatically build a BLFSW from possibilistic data.

3.1 Background on Possibility Theory and Defaults

In [15], possibility theory is introduced as a basis for qualitative decision theory. The author relate the expected pay-off $u(x)$ of a situation x to a preference relation $\preceq$ over situations s.t. $x \preceq y$ iff $u(x) \geq u(y)$. In presence of uncertainty, i.e., when situations are not precisely known, the belief state about what is the actual situation is represented by a possibility distribution π. The theory of possibility is a qualitative setting first introduced by Zadeh [16] and further developed by Dubois and Prade in [17]. It is qualitative in the sense that the only operations required are $\max$, $\min$ and order-reversing operations. However, numbers in the scale [0,1] are often used for convenience but the exact values of the numbers are not meaningful, it is only their order in the scale that is taken into account.

A possibility distribution π is used to compare the plausibility of situations: $\pi(x) \leq \pi(x')$ means that it is at least as plausible for x' to be the actual situation as for x to be it. $\pi(x) = 0$ means impossibility, $\pi(x) = 1$ means that x is unsurprising or normal. The state of total ignorance is represented by a possibility distribution where any situation is totally possible ($\forall x, \pi(x) = 1$). In order to reason on formula (hence sets of situations), two measures Π and N are defined: the possibility measure Π evaluates how unsurprising a formula is, hence $\Pi(\varphi) = 0$ means that φ is bound to be false. The necessity measure is its dual defined by $N(\varphi) = 1 - \Pi(\neg\varphi)$: $N(\varphi) = 1$ means that φ is bound to be true. N is defined from a possibility distribution π by: $N(\varphi) = min_{\omega \models \neg\varphi}(1 - \pi(\omega))$: a formula is all the more necessary as its counter models are less plausible.

In [4], the authors show that the utility of a decision d can be evaluated by combining the plausibilities $\pi(x)$ of the states x in which d is made and the utility $u(d(x))$ of the possible resulting state $d(x)$ after d, where $u(d(x))$ represents the satisfaction to be in the precise situation $d(x)$ (it is equal to the membership degree to the fuzzy set of preferred situations). The pessimistic criterion has been first introduced by Whalen [18] and leads to a pessimistic utility level of a decision d defined as follows: $u_{pes}(d) = \inf_{x \in X} \max(1 - \pi(x), u(d(x)))$. The optimistic criterion has been first proposed by Yager [19] and is defined by: $u_{op}(d) = \sup_{x \in X} \min(\pi(x), u(d(x)))$.

In possibilistic decision theory, the scales for possibilities and utilities are the same, hence, commensurable. In our proposal the commensurability of the two scales is not required: we do not aggregate possibilities and utilities, we rather use a kind of chance constrained approach [20, 21] in which they are dealt with separately.

Since a decision principle represents a defeasible reason to believe that some goal is achieved, we also need to recall some basics about handling defeasible rules in a possibilistic setting. A defeasible rule is a compact way to express a general rule without mentioning every exception to it. In a BLFSW the exceptions to a decision principle are its inhibitors. The conditional possibility measure denoted $\Pi(\varphi|\psi)$ is the possibility that φ holds in the worlds where ψ holds. It is related to the conditional possibility distribution as follows: $\Pi(\varphi|\psi) = \max_\omega \min(\Pi(\varphi|\omega), \pi(\omega|\psi))$.

A default rule $a \rightsquigarrow b$ translates, in the possibility theory framework, into the constraint $\Pi(a \wedge b) > \Pi(a \wedge \neg b)$ which expresses that having b true is strictly more possible than having it false when a is true [22]. Note that the constraint $\Pi(a \wedge b) > \Pi(a \wedge \neg b)$ is equivalent to $N(b|a) > 0$. Hence, if we know a and we search for a conclusion which satisfies the constraint $N() > 0$ then a solution is b. In this sense, decision principles are related to chance constraints in quantitative optimization problem. In this article, we will use the min conditioning ($|_{min}$) since we are interested in qualitative decision problems, i.e.,

$$\pi(\omega \,|_{min}\, \varphi) = \begin{cases} 1 & \text{if } \omega \models \varphi \text{ and } \pi(\omega) = \Pi(\varphi); \\ \pi(\omega) & \text{if } \omega \models \varphi \text{ and } \pi(\omega) < \Pi(\varphi); \\ 0 & \text{if } \omega \not\models \varphi \end{cases}$$

3.2 Interpreting a BLFSW in Possibility Theory

This section is devoted to give an interpretation of Support/Inhibitor and strength of a DP in a possibilistic setting. This will allow the designer of a Decision System to move from one formalism to another in order to check the accuracy of his proposed model. In addition, Possibility theory is recognized as a theory taking into account uncertainty and qualitative reasoning, so showing that there is a translation from a possibilistic representation of uncertainty and preferences to a BLFSW increases the validity of this framework. The BLFSW is able to take into account the degree of certainty of a DP which is not possible in a plain BLF. Nevertheless the possibilistic meaning of a DP and an inhibitor in a BLFSW are the same as those found for a BLF in [14]. First, we restate Π-DP and Π-inhibitor definitions: in order to be well-defined a DP has to be informative, i.e., the DP $\varphi \rightsquigarrow g$ is well-defined if the necessity of the goal g increases when φ holds:

Definition 8 (Π-DP [14])**.** *Given a possibility measure Π, a Π-DP $\varphi \rightsquigarrow g$ is s.t. $N(g|\varphi) > 0$*

In other words, the DP is the piece of knowledge which increases the certainty that the goal is realized. In the same way we can interpret the notion of inhibitor and support in possibility theory: an inhibitor ψ makes the default rule $\varphi \rightsquigarrow g$ no more valid in such a way that we are no longer sure that g will be realized when φ and ψ hold together. More precisely, ψ can be defined as an inhibitor of $\varphi \rightsquigarrow g$ if when ψ holds, the necessity of g being achieved (which was previously > 0) is reduced to zero.

Definition 9 (Π-Inhibitor [14])**.** *Given a possibility measure Π, the pair (ψ, p) is a Π-Inhibitor of the DP $p = \varphi \rightsquigarrow g$ if $N(g|\varphi \wedge \psi) = 0$*

In contrast, the support increases the certainty of the default rule. So when the support ψ holds, we are more sure that g will be realized.

Definition 10 (Π-Support)**.** *Given a possibility measure Π, the pair (ψ, p) is a Π-Support of the DP $p = \varphi \rightsquigarrow g$ if $N(g|\varphi \wedge \psi) > N(g|\varphi)$*

Moreover to complete the interpretation of a BLFSW in possibility theory we need to define the global strength $\alpha(p)$ of a DP p in possibilistic terms.

Definition 11 (Π-weight)**.** *Given a possibility measure* Π *and a weight function* w*.* w *is a* Π*-weight function iff for all possible* K*-BLFSW* ($\mathcal{P}_K$, $\mathcal{I}_K$, $\mathcal{S}_K$, *pol*, $\preceq$, w_K) *where* w_K *is the restriction of* w *on* $\mathcal{I}_K \cup \mathcal{S}_K$ *and for all decision principles* $p = \varphi \rightsquigarrow g$, $p' = \varphi' \rightsquigarrow g' \in \mathcal{P}_K$

- $\alpha(p) < 0$ *iff* $N(g \,|\, \varphi \bigwedge_{\psi \in \mathcal{I}_K(p) \cup \mathcal{S}_K(p)} \psi) = 0$
- $\alpha(p) \geq 0$ *iff* $N(g \,|\, \varphi \bigwedge_{\psi \in \mathcal{I}_K(p) \cup \mathcal{S}_K(p)} \psi) > 0$
- $\alpha(p) \geq \alpha(p') \geq 0$ *iff*
$N(g \,|\, \varphi \bigwedge_{\psi \in \mathcal{I}_K(p) \cup \mathcal{S}_K(p)} \psi) \geq N(g' \,|\, \varphi' \bigwedge_{\psi \in \mathcal{I}_K(p') \cup \mathcal{S}_K(p')} \psi) > 0$

with $\alpha(p)$ *defined from* w_K *according to Definition 3.*

In other words, the activation level $\alpha(p)$ of $p = \varphi \rightsquigarrow g$ defined in Definition 3 should reflect the certainty about the default rule $\varphi \rightsquigarrow g$ and should behave as stated in Definition 3: a negative activation level means that the default rule does not hold in presence of all its supports and inhibitors, a strictly positive one means that the goal is all the more likely to be achieved that the level is high, two distinct positive activation levels should be ranked according to the two necessities of the DPs. This last point will allow us to rank order alternatives more precisely. Using the definitions above we are now in position to define a Π-BLFSW.

Definition 12 (Π-BLFSW)**.** *A BLFSW* $B = (\mathcal{P}, \mathcal{I}, \mathcal{S}, pol, \preceq, w)$ *is a* Π*-BLFSW iff there exists a possibility distribution* π *over* Ω *and a utility function* u *on the set of goals* LIT_G*, s.t.*

- $\forall p = \varphi \rightsquigarrow g \in \mathcal{P}$, $u(g) \neq 0$
- $\forall g \in LIT_G$, $pol(g) = \oplus$ *iff* $u(g) > 0$
- $\forall g, g' \in LIT_G$, $g \preceq g'$ *iff* $u(g) \leq u(g')$
- *for all consistent knowledge base* K, $\forall g \in LIT_G$ *s.t.* $u(g) \neq 0$, $\forall \omega \in \Omega$, $\pi(g|\omega)$ *satisfies the constraints of Definitions 8, 9, 10 and 11*

Intuitively, in a Π-BLFSW the polarities and importances of the goals are based on a utility function and the weights on supports and inhibitors of DPs are consistent with the necessities of the default rules associated to DPs.

Thanks to this last definition, the designer can check whether her BLFSW is a Π-BLFSW hence whether it is consistent wrt a classical qualitative theory of uncertainty. If the possibility distribution does not seem realistic to the designer, she should modify the BLFSW (which summarizes it).

Remark 1. *It may happen that the agents want to distinguish between the strengths of two DPs* $p_1 = (\varphi_1 \rightsquigarrow g_1)$ *and* $p_2 = (\varphi_2 \rightsquigarrow g_2)$ *because she knows that* $N(g_1|\varphi_1) > N(g_2|\varphi_2)$*. In order to do that, she may use the notion of support, by adding a support* $s_1 = (\varphi_1, p_1)$*. In that case,* $\alpha_{p_1} = w(s_1)$ *is necessarily greater than* $\alpha_{p_2} = 0$*.*

The following proposition shows the relation between the weight associated to a goal in a K-BLFSW and its necessity to hold wrt this knowledge base K. The ranking on alternatives described by K based on the goals achieved in K is the same as the one obtained in a Π-BLFSW based on K.[2]

Proposition 1. *Given a* Π*-BLFSW* $B = (\mathcal{P}, \mathcal{I}, \mathcal{S}, pol, \preceq, w)$ *built on a possibility distribution* π *on the set of worlds* Ω *and on a utility function* u *on* LIT_G*.* B *is s.t. for all consistent knowledge bases* K, K' *and for any goals* g, g'*:*

- g *not realized wrt* B_K *iff* $N(g|K) = 0$ *or* $u(g) = 0$
- *if* g *realized wrt* B_K *and* g' *realized wrt* $B_{K'}$ *then* $\alpha_K(g) > \alpha_{K'}(g')$ *iff* $N(g|K) > N(g'|K')$

Proof. (sketch) The first item follows from definitions 8 and 12, the second one from definitions 3 and 11. □

3.3 From Possibility theory to BLFSW: An Example

Building a BLFSW can be done in two independent steps: first define the DPs and their weighted relationships from a given possibility distribution, second use the utilities of goals in order to filter out DPs with goals of null utility and to rank the DPs in the BLFSW. Here, we only present the first step based on the knowledge of a possibility distribution over all possible worlds $\omega \in \Omega$ and the possibility for each goal to be true in each world (Table.1).

For each goal in LIT_G (here in $\{s, \neg s, c, \neg c, e, \neg e\}$), we check whether we can generate a DP concluding it by checking Definition 8, on all the conjunctive formulas that can be built, starting from formulas restricted to a single literal and adding new literals progressively. Let us consider the goal s when we know p, we have $N(s|p) = 1 - \Pi(\neg s|p) = 1 - \max_\omega(\min(\pi(\omega|p), \Pi(\neg s|\omega))) = 1 - 0.4 = 0.6 > 0$ hence $p_1 = p \rightsquigarrow s$ is a DP. If we suppose that we know w, $N(s|w) = 1 - \Pi(\neg s|w) = 1 - 1 = 0$ due to the world ω_{12} hence $s \rightsquigarrow w$ is not a DP. Let us look for supports and inhibitors, we have $N(s|p \wedge f) = 0.8 > N(s|p)$, so due to definition 10, $s_1 = (f, p_1)$ is a support. p_1 has also an inhibitor since adding $\neg w$ we get $N(s|p \wedge \neg w) = 0$. $N(s|p \wedge f \wedge \neg w) = 0.7$ thus $s_2 = (f \wedge \neg w, p_1)$ is also a support. Using the same process on the other two goals, we obtain $p_2 = w \rightsquigarrow c$, $N(c|w) = 0.6$, $i_2 = (f, p_2)$, $N(c|w \wedge f \wedge o) = 0$, $s_3 = (o, p_2)$, $N(c|w \wedge o) = 0.7$, $i_3 = (f \wedge o, p_2)$, $N(c|w \wedge f \wedge o) = 0$, $p_3 = (f, e)$, $N(e|f) = 0.6$ and $i_4 = (o, p_3)$,

[2] This ranking can be obtained with the relations $\preceq_{BiLexi}$ or $\preceq_{BW}$.

ω	$\pi(\omega)$	$\Pi(s\|\omega)$	$\Pi(\neg s\|\omega)$	$\Pi(c\|\omega)$	$\Pi(\neg c\|\omega)$	$\Pi(e\|\omega)$	$\Pi(\neg e\|\omega)$
ω_1: $pwfo$	0.3	1	0	1	1	0.2	1
ω_2: $pwf\neg o$	0.3	1	0	0.8	1	1	0.4
ω_3: $pw\neg fo$	1	1	0	1	0.3	0	1
ω_4: $pw\neg f\neg o$	1	1	0	1	0.4	0	1
ω_5: $p\neg wfo$	0.2	1	0.3	0.2	1	0.2	1
ω_6: $p\neg wf\neg o$	0.2	1	0.3	0.1	1	1	0.4
ω_7: $p\neg w\neg fo$	0.4	1	1	0.8	1	0	1
ω_8: $p\neg w\neg f\neg o$	0.4	1	1	0	1	0	1
ω_9: $\neg pwfo$	0.3	0	1	1	1	0	1
ω_{10}: $\neg pwf\neg o$	0.3	0	1	0.8	1	1	0.4
ω_{11}: $\neg pw\neg fo$	1	0	1	1	0.3	0	1
ω_{12}: $\neg pw\neg f\neg o$	1	0	1	1	0.4	0	1
ω_{13}: $\neg p\neg wfo$	0.3	0	1	0.2	1	0	1
ω_{14}: $\neg p\neg wf\neg o$	1	0	1	0.1	1	1	0.4
ω_{15}: $\neg p\neg w\neg fo$	1	0	1	0.2	1	0	1
ω_{16}: $\neg p\neg w\neg f\neg o$	0.4	0	1	0	1	0	1

Table 1: Possibility distributions on worlds and goals

$N(e|f \wedge o) = 0$. Let us now focus on the weight assignments. The weights must satisfy all the constraints entailed by Definition 11. For instance, $w_{s_1} > w_{s_1} + w_{s_2} - w_{i_1} > 0$ and $0 \geq -w_{i_1}$. Note that if $N(s|p \wedge f \wedge \neg w) = N(s|p \wedge f) = 0.6$ then $w(f \wedge \neg w, P_1) = w(s|p \wedge \neg w)$. In that case the inhibitor $\neg w$ is cancelled by the support $f \wedge \neg w$. The possible assignments of weights are infinite, for instance the one given in Example 1: $w_{p_1} = 0.4$, $w_{s_1} = 0.6$, $w_{s_2} = 0$, $w_{i_1} = 0.4$ satisfies the constraints. Using the same process on the other two goals, for c we have $w_{p_2} + w_{s_3} - w_{i_3} \leq 0$, $w_{p_2} - w_{i_3} \leq 0$ and $w_{s_3} > 0$ for instance the one given in Example 1: $w_{i_3} = 0.9$, $w_{s_3} = w_{p_2} = 0.1$. So the decision problem defined by Table.1 is equivalent to the BLFSW of Example 1.

4 Conclusion

This paper proposes an extension of the BLF of [1] in order to deal with supports and weights.The framework is a visual way to encode and explain a qualitative decision theory. The BLFSW's main benefit is to provide a representation that allows a user to clearly express the principles and utility levels which govern the decision process. In BLFSW, the decision is justified by the importance and polarities of the tangible results that are realized if the alternative is chosen, these results are also explained by the valid principles (not inhibited DPs) that apply in the situation. The ability to explain how the weights of inhibitors and supports of DPs are computed is one of the main result of this paper. This result is based on a procedure that

builds a BLFSW from utilities and uncertain knowledge expressed in possibilistic terms.

References

[1] F. Dupin De Saint Cyr and R. Guillaume, "Group Decision Making in a Bipolar Leveled Framework," in *PRIMA*, 2017, pp. 34–52.

[2] A. Tchangani, Y. Bouzarour-Amokrane, and F. Pérès, "Evaluation model in decision analysis: Bipolar approach," *Informatica*, vol. 23, no. 3, pp. 461–485, 2012.

[3] H. Raiffa, *Decision Analysis: Introductory lectures on choices under uncertainty*. Reading, Massachusetts: Addison-Wesley, 1970.

[4] D. Dubois, H. Prade, and R. Sabbadin, "Decision-theoretic foundations of qualitative possibility theory," *Euro. J. of Operational Research*, vol. 128, no. 3, pp. 459–478, 2001.

[5] D. Dubois and H. Fargier, "Qualitative bipolar decision rules: Toward more expressive settings," in *Preferences and Decisions*. Springer, 2010, pp. 139–158.

[6] M. J. Wooldridge, *Reasoning about rational agents*. MIT press, 2000.

[7] L. Amgoud, C. Devred, and M.-C. Lagasquie-Schiex, "A constrained argumentation system for practical reasoning," in *Argumentation in Multi Agent Systems*. Springer, 2009, pp. 37–56.

[8] J. Raz, *Practical reasoning*. Oxford University Press, 1978.

[9] D. Walton, *Argumentation schemes for presumptive reasoning*. (first edition 1996) Routledge, 2013.

[10] L. Amgoud and H. Prade, "Comparing decisions on the basis of a bipolar typology of arguments," in *Preferences and Similarities*, G. Della Riccia, D. Dubois, R. Kruse, and H. J. Lenz, Eds. Springer, 2008, pp. 249–264.

[11] L. Amgoud and S. Vesic, "A formal analysis of the role of argumentation in negotiation dialogues," *Journal of Logic and Computation*, vol. 22, pp. 957–978, 2012.

[12] I. Rahwan, S. D. Ramchurn, N. R. Jennings, P. McBurney, S. Parsons, and L. Sonenberg, "Argumentation-based negotiation," *Knowledge Engineering Review*, vol. 18, no. 4, pp. 343–375, 2003.

[13] J. Bonnefon, D. Dubois, and H. Fargier, "An overview of bipolar qualitative decision rules," in *Preferences and Similarities*, ser. vol 504, CISM. Springer, 2008, pp. 47–73.

[14] F. Dupin De Saint Cyr and R. Guillaume, "Analyzing a Bipolar Decision Structure through Qualitative Decision Theory," *KI - Künstliche Intelligenz*, vol. 31, no. 1, pp. 53–62, mars 2017.

[15] D. Dubois and H. Prade, "Possibility theory: qualitative and quantitative aspects," in *Quantified Representation of Uncertainty and Imprecision*. Kluwer Academic Publishers, 1998, pp. 169–226.

[16] L. Zadeh, *Fuzzy Sets as a Basis for a Theory of Possibility*, ser. Memorandum: Electronics Research Laboratory. U. of California, 1977.

[17] D. Dubois and H. Prade, *Possibility theory*. Plenum Press, New York, 1988.

[18] T. Whalen, "Decision making under uncertainty with various assumptions about available information," *IEEE Trans. on Systems, Man and Cybern.*, vol. 14, no. 6, pp. 888–900, 1984.

[19] R. Yager, "Possibilistic decision making," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, pp. 388–392, 1979.

[20] A. Charnes and W. Cooper, "Chance-constrained programming," *Management science*, vol. 6, no. 1, pp. 73–79, 1959.

[21] P. Li, H. Arellano-Garcia, and G. Wozny, "Chance constrained programming approach to process optimization under uncertainty," *Computers & chemical engineering*, vol. 32, no. 1-2, pp. 25–45, 2008.

[22] S. Benferhat, D. Dubois, and H. Prade, "Representing default rules in possibilistic logic," in *KR*, 1992, pp. 673–684.