

A Unified Argumentation-Based Framework for Knowledge Qualification

Loizos Michael

Open University of Cyprus
loizos@ouc.ac.cy

Antonis Kakas

University of Cyprus
antonis@cs.ucy.ac.cy

Abstract

Among the issues faced by an intelligent agent, central is that of reconciling the, often contradictory, pieces of knowledge — be those given, learned, or sensed — at its disposal. This problem, known as *knowledge qualification*, requires that pieces of knowledge deemed reliable in some context be given preference over the others.

These preferences are typically viewed as encodings of *reasoning patterns*; so, the frame axiom can be encoded as a preference of persistence over spontaneous change. Qualification, then, results by the principled application of these preferences. We illustrate how this can be naturally done through argumentation, by uniformly treating object-level knowledge and reasoning patterns alike as arguments that can be defeated by other stronger ones. We formulate an argumentation framework for Reasoning about Actions and Change that gives a semantics for Action Theories that include a State Default Theory.

Due to their explicit encoding as preferences, reasoning patterns can be adapted, when and if needed, by a domain designer to suit a specific application domain. Furthermore, the *reasoning patterns can be defeated* in lieu of stronger external evidence, allowing, for instance, the frame axiom to be overridden when unexpected sensory information suggests that spontaneous change may have broken persistence in a particular situation.

Introduction

When reasoning, an agent is often required to take into account numerous pieces of static and temporal knowledge at its disposal, whose implications are not always consistent. In this setting, the various pieces of knowledge need to qualify each other, and the precise manner in which this should happen is known as the *knowledge qualification* problem.

In addition to the well-studied frame, ramification, and action qualification problems, knowledge qualification encompasses also the recently proposed problem of *integrating default static and default temporal reasoning* [Kakas, Michael, and Miller, 2008a]. Motivated by this problem, and following our past work on it [Kakas, Michael, and Miller, 2008b; Michael and Kakas, 2009], we address it herein by tackling the general knowledge qualification problem directly.

We follow an argumentation-based approach. Each piece of knowledge is encoded as an argument in support of some

conclusion. The qualification comes down to imposing preferences among the arguments and allowing stronger arguments to defeat weaker ones. These preferences correspond in a very precise sense to the reasoning patterns one uses to draw conclusions. The frame axiom, for instance, can be encoded as follows: an argument stating that properties persist; an argument stating that properties change spontaneously; a preference of the former over the latter; and preferences of causal changes known in the domain over persistence.

Instead of encoding such preferences implicitly in the semantics, we include them explicitly in the domain description. Thus, both object-level knowledge and reasoning patterns are uniformly treated as arguments, which can be defeated in the presence of stronger arguments. It is thus possible, for a domain designer to choose to override some of the typical reasoning patterns in a particular domain, by including the inverse preferences and making the latter stronger.

The same preference reversal can be used also in scenarios where an agent's conclusions contradict its sensory inputs. The unexpected sensory information can be accommodated simply by introducing exogenous arguments (not supported directly by the knowledge at the agent's disposal), and use them to defeat the endogenous arguments (both object-level and reasoning patterns) that lead to the particular contradiction. In such cases, we have a phenomenon known as *exogenous* qualification, which is to be distinguished from *endogenous* qualification, where the applicable arguments are supported directly by the knowledge available to the agent.

Indeed, our argumentation-based approach not only provides a uniform solution to both endogenous and exogenous qualification, but it makes their distinction clear. We review the proposed framework's formal semantics in the next section, and illustrate its main features through an in-length discussion of the Fred meets Tweety domain [Kakas, Michael, and Miller, 2008a] in the section thereafter.

An Argumentation-Based Semantics

A Typical Argumentation Framework

An argumentation framework in its general abstract form [Dung, 1995] is defined in terms of a set of arguments and an attacking relation between the arguments. Its semantics is usually based on the notion of *admissible* argument subsets and refinements of this notion. Admissible argument subsets

are those that are not self-attacking, and attack all argument subsets that attack them; i.e., their counter-arguments.

A common way to realize an argumentation framework, linking the arguments and the attacking relation to the application domain information, is to use *preferences or priorities* between arguments to give a notion of relative strength between them [Prakken and Sartor, 1997; Kakas and Moraitis, 2003; Dimopoulos, Moraitis, and Amgoud, 2008], and realize, thus, the generally non-symmetric attacking relation.

In many cases, as shown originally in [Prakken and Sartor, 1997], the priority between arguments becomes part of the argumentation framework, with arguments for the priorities themselves, and arguments for (higher-order) priorities between the different priority arguments. The same argumentation process that is used for the object-level arguments can be applied to the priority arguments that refer to the former, and iteratively so for the higher-order arguments involved.

We will work within such a preference-based argumentation framework where we have an argumentation language \mathcal{L} that contains a clause for every possible argument, including clauses of the form $\alpha_1 \succ_j \alpha_2$ indicating that argument α_1 is preferred to argument α_2 according to the preference relation \succ_j . Associated with \mathcal{L} is an entailment operator \models . The definitions follow those in typical preference-based argumentation frameworks, with the generalization of allowing multiple preference relations; we discuss this feature later.

Definition 1 (Strength) An argument set $\mathcal{A}_1 \subseteq \mathcal{L}$ is *weaker* than an argument set $\mathcal{A}_2 \subseteq \mathcal{L}$ **given** some background theory \mathcal{B} *under* a preference relation \succ_j if there exist arguments $\alpha_1^1 \in \mathcal{A}_1$, $\alpha_2^1 \in \mathcal{A}_2$ such that $\mathcal{A}_2 \cup \mathcal{B} \models \alpha_2^1 \succ_j \alpha_1^1$ and there exist no arguments $\alpha_1^2 \in \mathcal{A}_1$, $\alpha_2^2 \in \mathcal{A}_2$ such that $\mathcal{A}_1 \cup \mathcal{B} \models \alpha_1^2 \succ_j \alpha_2^2$.

Definition 2 (Attack) An argument set $\mathcal{A}_1 \subseteq \mathcal{L}$ **attacks** an argument set $\mathcal{A}_2 \subseteq \mathcal{L}$ **given** some background theory \mathcal{B} *under* a preference relation \succ_j if there exist subsets $\mathcal{A}_1^m \subseteq \mathcal{A}_1$, $\mathcal{A}_2^m \subseteq \mathcal{A}_2$ such that the following conditions hold:

- (i) $\mathcal{A}_1^m \cup \mathcal{B} \models \phi$ and $\mathcal{A}_2^m \cup \mathcal{B} \models \neg\phi$ for some clause $\phi \in \mathcal{L}$;
- (ii) no strict subsets of \mathcal{A}_1^m and \mathcal{A}_2^m satisfy Condition (i);
- (iii) \mathcal{A}_1^m is not weaker than \mathcal{A}_2^m .

Definition 3 (Admissibility) An argument set $\mathcal{A} \subseteq \mathcal{L}$ is *j-admissible* **given** a background theory \mathcal{B} if the following conditions hold:

- (i) \mathcal{A} does not attack \mathcal{A} given \mathcal{B} under \succ_j ;
- (ii) for every argument set $\mathcal{A}' \subseteq \mathcal{L}$ that attacks \mathcal{A} given \mathcal{B} under \succ_j , \mathcal{A} attacks \mathcal{A}' given \mathcal{B} under \succ_j .

Argumentation for Knowledge Qualification

We now develop an argumentation framework for Reasoning about Actions and Change, and show how it gives a semantics for Action Theories that include a State Default Theory.

For simplicity, and as it suffices for our purposes, we shall henceforth assume \mathcal{F} to be a set of propositions (i.e., properties of interest), and \mathcal{T} to be the set of non-negative integers. Whenever predicates are used instead of propositions, they are used as shorthand for the set of all the corresponding ground predicates over a fixed set of constants implied by the

context. We, now, instantiate the argumentation framework presented earlier. Below $\text{form}[\Sigma]$ is the set of propositional formulas over alphabet Σ with the usual connectives.

Definition 4 (Argumentation Language) The *argumentation language* is defined to be $\mathcal{L} \triangleq \text{form}[\bigcup_{k=0}^{\infty} \mathcal{F}_k]$, where $\mathcal{F}_0 \triangleq \mathcal{F} \times \mathcal{T}$, and for every integer $k \geq 1$, \mathcal{F}_k contains all arguments $(\alpha_1 \succ_j \alpha_2)$, where $\alpha_1, \alpha_2 \in \text{form}[\bigcup_{m=0}^{k-1} \mathcal{F}_m]$.

The atomic base arguments F at $T \in \mathcal{F}_0$ stipulate that property F holds at time-point T . Possible arguments of this form include *alive(tweety)* at 3 and *shoot(tweety)* at 4.

Composite base arguments are formulas over the atomic base arguments. For any time-point $T \in \mathcal{T}$ and constant X ,

$$\begin{aligned} & \text{alive}(X) \text{ at } T-1 \rightarrow \text{alive}(X) \text{ at } T \\ & \neg(\text{bird}(X) \text{ at } T \wedge \text{flying}(X) \text{ at } T \wedge \neg(\text{alive}(X) \text{ at } T)) \\ & \text{shoot}(X) \text{ at } T-1 \wedge \text{loaded} \text{ at } T-1 \rightarrow \neg(\text{alive}(X) \text{ at } T) \end{aligned}$$

state that: (i) if X is alive, it remains so; (ii) a dead bird does not fly; and (iii) if X is shot with a loaded gun, it is killed.

For readability, we shall not distribute the time-point over each proposition, and we shall use curly brackets to group propositions at the same time-point. The above become

$$\begin{aligned} & \{\text{alive}(X)\} \text{ at } T-1 \rightarrow \{\text{alive}(X)\} \text{ at } T \\ & \{\neg(\text{bird}(X) \wedge \text{flying}(X) \wedge \neg\text{alive}(X))\} \text{ at } T \\ & \{\text{shoot}(X) \wedge \text{loaded}\} \text{ at } T-1 \rightarrow \{\neg\text{alive}(X)\} \text{ at } T \end{aligned}$$

Higher-order arguments postulate preferences over lower-order arguments. To state that the causal change (clause 3) from above takes precedence over the frame axiom (clause 1) and the static constraint (clause 2) from above we write

$$\begin{aligned} & \{\text{shoot}(X) \wedge \text{loaded}\} \text{ at } T-1 \rightarrow \{\neg\text{alive}(X)\} \text{ at } T \\ & \succ_1 \{\text{alive}(X)\} \text{ at } T-1 \rightarrow \{\text{alive}(X)\} \text{ at } T \\ & \{\text{shoot}(X) \wedge \text{loaded}\} \text{ at } T-1 \rightarrow \{\neg\text{alive}(X)\} \text{ at } T \\ & \succ_1 \{\neg(\text{bird}(X) \wedge \text{flying}(X) \wedge \neg\text{alive}(X))\} \text{ at } T \end{aligned}$$

For readability, we can alternatively write the preferences $3(X,T) \succ_1 1(X,T)$ and $3(X,T) \succ_1 2(X,T)$, making explicit in the clause number the instantiation of the variables involved.

Such preferences will be used to encode typical reasoning patterns; e.g., causal change is preferred over persistence, which is preferred over static knowledge, which is preferred over causal change [Kakas, Michael, and Miller, 2008a; Michael and Kakas, 2009]. In situations where these typical reasoning patterns are not applicable (e.g., “strong” actions that override the static theory), the domain designer may encode these exceptions by including the inverse preference, and then including a meta-preference stating the conditions under which the exceptional preference should override the typical preference; the next section illustrates this scenario.

Definition 5 (Basic Constructs) *Entailment* for \mathcal{L} is defined in the typical manner. A *state* \mathcal{S} (at T) is a set of literals over $\mathcal{F} \times \{T\}$ for some $T \in \mathcal{T}$ and so that each $F \in \mathcal{F}$ appears uniquely. An *argument set* is a subset $\mathcal{A} \subseteq \mathcal{L}$. A *narrative* is a subset $\mathcal{N} \subseteq \text{form}[\mathcal{F} \times \mathcal{T}]$. A *domain* \mathcal{D} is a mapping from time-points T to subsets $\mathcal{D}_T \subseteq \mathcal{L}$.

States capture what holds at a time-point, and their literals persist and change following the reasoning patterns encoded

as preference arguments. We take the view that only object-level information is recorded in states. In particular, this implies that preferences encoding reasoning patterns will not be subject to persistence and change by the explicit invocation of actions — instead, they hold everywhere by default.¹

With regards to the semantics of the proposed framework, we advocate a pragmatic *online* (to the extent possible) approach. Starting with $T = 0$, the current state $\mathcal{M}(T-1)$ acts as a background theory, and the arguments in \mathcal{D}_T (those encoding object-level knowledge and reasoning patterns alike) are used to reason about what holds in the next state $\mathcal{M}(T)$ at time-point T . The agent is expected to construct an admissible argument set $\theta(T)$ of *endogenous* arguments in \mathcal{D}_T , so that $\mathcal{M}(T)$ will respect the constraints of the narrative \mathcal{N} .²

In certain cases no use of endogenous arguments in \mathcal{D}_T can account for the narrative (e.g., in case of an unexpected observation following an action occurrence). In those cases we allow $\theta(T)$ to use *exogenous* arguments outside \mathcal{D}_T , appealing to reasons (be those processes, events, constraints, or even reasoning patterns) outside the agent’s theory of its environment. Use of such exogenous arguments is minimized, giving preference to arguments supported by the agent’s theory. Besides this minimization, however, the argumentation framework offers a uniform treatment of endogenous and exogenous arguments, while making precise their distinction.

Definition 6 (Interpretation) An *interpretation* \mathcal{M} of a domain \mathcal{D} is a set $\bigcup_{T \in \mathcal{T}} \mathcal{M}(T)$, where $\mathcal{M}(T)$ is a state at T ; let $\mathcal{M}(-1) \triangleq \emptyset$. Given an interpretation \mathcal{M} of \mathcal{D} , a mapping $\theta : \mathcal{T} \rightarrow 2^{\mathcal{L}}$, and a narrative \mathcal{N} , we shall say that: \mathcal{M} *j -accepts* θ if $\theta(T)$ is j -admissible given $\mathcal{M}(T-1)$, for every $T \in \mathcal{T}$; \mathcal{M} *agrees with* θ if $\theta(T) \cup \mathcal{M}(T-1) \models \mathcal{M}(T)$, for every $T \in \mathcal{T}$; \mathcal{M} *accounts for* \mathcal{N} if $\mathcal{M} \models \mathcal{N}$.

An interpretation \mathcal{M} is a sequence of states, for which we wish to identify a sequence θ of argument sets that explain the state transitions in \mathcal{M} . This is achieved by having each $\theta(T)$ be admissible given the state at $T-1$ (i.e., 1-accepted) and correctly predict the state at T (i.e., agreed with). In addition, we expect that the narrative³ is accounted for by \mathcal{M} , and that the use of exogenous arguments is minimized.

¹This view implies that if a given reasoning pattern is violated (for any reason) at some time-point (e.g., an action’s effect is not brought about), then this violation is local and does not persist (e.g., action invocations are oblivious to their earlier failures). States can, of course, be trivially expanded to include the reasoning patterns as well. This would allow to reason about domains in which an action occurrence has as an effect that some reasoning pattern comes into (dis)use: e.g., flipping the fan switch disengages the frame axiom from applying to the fan position; or, the failure of opening a door will persist and imply that all future attempts for opening the door will fail. We do not discuss such scenarios further in this work.

²We treat action occurrences and fact observations uniformly as constraints in a narrative. In the spirit of our approach, the distinction that action occurrences do not typically persist across time, like fact observations, is captured by explicitly including this reasoning pattern as an argument; the next section illustrates this scenario.

³A special category is that of online narratives, where each constraint references only one time-point. In such a case, the requirement that narratives be accounted for need not be imposed globally and a posteriori, as done in Definition 7. Instead, it can be con-

Definition 7 (Model for Narrative) A *model* \mathcal{M} of a domain \mathcal{D} for a narrative \mathcal{N} is an interpretation of \mathcal{D} that accounts for \mathcal{N} , and 1-accepts and agrees with a mapping θ such that no interpretation \mathcal{M}' of \mathcal{D} that accounts for \mathcal{N} , and 1-accepts and agrees with a mapping θ' , is such that $\bigcup_{T \in \mathcal{T}} (\theta'(T) \setminus \mathcal{D}_T) \subset \bigcup_{T \in \mathcal{T}} (\theta(T) \setminus \mathcal{D}_T)$.

Definition 7 is the first place in the exposition of the formal semantics where we fix the use of a preference relation. In particular, the *local* admissibility of each $\theta(T)$ is with respect to the preference relation \succ_1 only. The next definition imposes a second, *global* admissibility requirement over the entire θ with respect to the preference relation \succ_2 only.

Definition 8 (Preferred Model for Narrative) A *preferred model* \mathcal{M} of a domain \mathcal{D} for a narrative \mathcal{N} is a model of \mathcal{D} for \mathcal{N} that 1-accepts and agrees with a mapping θ that is 2-admissible given \emptyset .

Prediction and Explanation

The argumentation language allows for multiple preference relations. Our semantics makes use of exactly two such preference relations, \succ_1 and \succ_2 , which, roughly speaking, encode preferences between pairs of arguments that reference, respectively, the same time-point, or the entire time-line. This distinction is evident, after all, in Definitions 7 and 8.

Beyond this distinction, there is a deeper interpretation of the use of two preference relations. One encodes preferences among arguments in terms of their *predictive* power, while the other encodes preferences among arguments in terms of their *explanatory* power. In fact, it is possible for a pair of arguments to have opposite preferences on these two axes.

In a real-world setting, for instance, observing something to hold offers, often, an indisputable argument that indeed it holds. Observations are, in this sense, preferred arguments for prediction. On the other hand, attempting to explain why things are the way they are by claiming that it is because they were observed to be so, seems to be a rather weak argument; a stronger one would be to argue that they were caused by some action, or persisted from the past. Observations are, in this sense, non-preferred arguments for explanation.⁴

Acknowledging the predictive and explanatory natures of arguments, Definition 7 identifies models that are highly preferred in terms of their predictive power, while among those, Definition 8 identifies models that are highly preferred in terms of their explanatory power.

As we shall illustrate in the next section, predictive preferences can be used to encode the (predictive) reasoning patterns of an agent (e.g., concluding that a causal effect should

be considered during the step-by-step argumentation performed by the agent, by including the constraints as strong arguments that attack all other arguments. However, to properly capture the constraint nature of the narrative, these arguments must be allowed only to attack but not defend admissibility; such an asymmetrical treatment of arguments is considered in [Kakas, Miller, and Toni, 1999].

⁴This precise nature of observations is what guided us to treat narrative as a distinct part of a domain that can filter which interpretations are models (i.e., predictions coming through observations should be respected), but cannot be used as an argument to construct such models (i.e., explanations coming through observations should be avoided). See also the preceding footnote.

override the persistence of its negation), while explanatory preferences can be used to encode a failure recovery policy in case exogenous arguments need to be employed to account for the narrative. Different recovery policies (e.g., concluding that some action must have failed to produce its effects, or that some property spontaneously stopped persisting) are available, and depending on the application domain one may wish to consider some particular subset of them.⁵

An Illustrative Example Domain

Consider a variant of the Fred meets Tweety domain [Kakas, Michael, and Miller, 2008a]:

Birds can, generally, fly. Penguins and turkeys are birds but cannot fly, with the rare exception of being magic as a result of a spell, in which case it is unclear if they can fly. Shooting someone causes noise, and kills that someone. Noise is heard only briefly, and causes birds to fly. Initially, the turkey Fred is alive, the bird Tweety is close by, and a gun is loaded. Fred is shot with that gun, and some time later Tweety is observed not to fly. What can be deduced about whether Fred is alive?

We incrementally build a representation \mathcal{D} of this domain, while discussing, as we progress, its possible models.

Initial Domain Representation

Although our framework treats domains as black boxes, and does not, hence, define a language for representing domains, it is straightforward to devise one such language by building on the argumentation language. Below, variable X is taken to range over the set $\{\text{fred, tweety}\}$. Having done this instantiation, we let \mathcal{F} contain all propositions in the resulting representation. In this representation, L is taken to range over the set of all literals over $\mathcal{F} \setminus \{\text{shoot}(X), \text{spell}(X)\}$, and A is taken to range over the set $\{\text{shoot}(X), \text{spell}(X)\}$. Variable T is treated as a formal parameter in the domain language. When the domain \mathcal{D} is mapped to an argument set \mathcal{D}_{T_0} at a specific time-point T_0 , as required by Definition 5, T is instantiated to the value T_0 , giving rise to a finite set of arguments in the (propositional) argumentation language.

The domain static knowledge is shown below.⁶

$$\begin{aligned}
S1(X,T) & : \{\text{penguin}(X) \vee \text{turkey}(X) \rightarrow \text{bird}(X)\} \text{ at } T \\
S2(X,T) & : \{\text{penguin}(X) \vee \text{turkey}(X) \rightarrow \neg \text{canfly}(X)\} \text{ at } T \\
S3(X,T) & : \{\text{bird}(X) \rightarrow \text{canfly}(X)\} \text{ at } T \\
S4(X,T) & : \{\neg \text{magic}(X)\} \text{ at } T \succ_1 S3(X,T) \\
S5(X,T) & : \{\neg \text{alive}(X) \rightarrow \neg \text{canfly}(X)\} \text{ at } T \\
S6(X,T) & : S5(X,T) \succ_1 S3(X,T) \\
S7(X,T) & : \{\neg \text{canfly}(X) \rightarrow \neg \text{flying}(X)\} \text{ at } T \\
S8(X,T) & : \{\neg \text{magic}(X)\} \text{ at } T
\end{aligned}$$

The domain causal knowledge is shown below.

$$\begin{aligned}
C1(X,T) & : \{\text{shoot}(X) \wedge \text{loaded}\} \text{ at } T-1 \rightarrow \{\text{noise}\} \text{ at } T \\
C2(X,T) & : \{\text{shoot}(X) \wedge \text{loaded}\} \text{ at } T-1 \rightarrow \{\neg \text{alive}(X)\} \text{ at } T \\
C3(X,T) & : \{\text{noise} \wedge \text{bird}(X)\} \text{ at } T-1 \rightarrow \{\text{flying}(X)\} \text{ at } T \\
C4(X,T) & : \{\text{spell}(X)\} \text{ at } T-1 \rightarrow \{\text{magic}(X)\} \text{ at } T
\end{aligned}$$

Implicit in the natural language description of the domain are typical reasoning patterns under which object-level knowledge is to be reasoned with. We discuss them next.

Consider, first, the usual frame axiom, stating that properties persist across time (unless, of course, explicitly caused otherwise), and the no-action axiom, stating that actions do not occur spontaneously. We, thus, include the arguments below (not just in this, but in all domain descriptions).

$$\begin{aligned}
FA(L,T) & : \{L\} \text{ at } T-1 \rightarrow \{L\} \text{ at } T \\
NA(A,T) & : \{\neg A\} \text{ at } T
\end{aligned}$$

Following the approach in [Kakas, Michael, and Miller, 2008a; Michael and Kakas, 2009], persistence is less preferred than explicit causal change, the latter is less preferred than the static knowledge, and the latter is less preferred than persistence. We, thus, include the preferences below, for each choice of the clauses $Si(X,T)$ and $Cj(X,T)$.

$$\begin{aligned}
C1FA(X,T) & : C1(X,T) \succ_1 FA(\neg \text{noise}, T) \\
C2FA(X,T) & : C2(X,T) \succ_1 FA(\text{alive}(X), T) \\
C3FA(X,T) & : C3(X,T) \succ_1 FA(\neg \text{flying}(X), T) \\
C4FA(X,T) & : C4(X,T) \succ_1 FA(\neg \text{magic}(X), T)
\end{aligned}$$

$$\begin{aligned}
SiCj(X,T) & : Si(X,T) \succ_1 Cj(X,T) \\
FASi(L,X,T) & : FA(L,T) \succ_1 Si(X,T)
\end{aligned}$$

The natural language description of the domain suggests two (domain-specific) exceptions to the typical (domain-independent) reasoning patterns in the preferences above.

The first exception is that *noise* is not bound by persistence (although $\neg \text{noise}$ is). This exception is accommodated by introducing an argument terminating noise that is more preferred than its persistence but less so than its causation. We, thus, allow noise to be caused momentarily to hold.

$$\begin{aligned}
T1(T) & : \{\text{noise}\} \text{ at } T-1 \rightarrow \{\neg \text{noise}\} \text{ at } T \\
T1FA(T) & : T1(T) \succ_1 FA(\text{noise}, T) \\
CIT1(X,T) & : C1(X,T) \succ_1 T1(T)
\end{aligned}$$

The second exception is that a spell will exempt someone from the default state of not being magic. In the terminology used in [Michael and Kakas, 2009], this is a “strong” action. This particular exception could be accommodated directly by simply flipping the preference in the clause $S8C4(X,T)$. An alternative, and what we suggest, is to keep the typical (domain-independent) preference of static knowledge over causal change unaffected, and introduce the flipped preference as an exception, introducing also a meta-preference giving priority to the domain-specific exceptional preference over the general, and domain-independent preference.⁷

⁵Of course, explanatory preferences can be used also when no failure needs to be recovered from, imposing, for instance a preference that, all things being equal, a persistence argument for *loaded* is preferred as an *explanation* than a causal argument for *loaded*.

⁶In terms of expressivity, note the use of preference $S4(X,T)$, stating that the preference applies only under certain conditions.

⁷In terms of expressivity, note the use of preference $C4S8P(X,T)$ over preferences $C4S8(X,T)$ and $S8C4(X,T)$, stating, roughly, that if one of the two lower-ordered preferences were to be violated, then it is preferred that the latter (domain-independent) one will be

$$C4S8(X,T) : C4(X,T) \succ_1 S8(X,T)$$

$$C4S8P(X,T) : C4S8(X,T) \succ_1 S8C4(X,T)$$

The narrative \mathcal{N} associated with the domain \mathcal{D} includes: $\{alive(fred)\}$ at 1, $\{turkey(fred)\}$ at 1, $\{bird(tweety)\}$ at 1, $\{loaded\}$ at 1, $\{shoot(fred)\}$ at 2, $\{\neg flying(tweety)\}$ at 5.

Typical Actions and Change

Consider an agent that attempts to build a model for domain \mathcal{D} . Let us assume first that the narrative \mathcal{N}' available to the agent is the one resulting from the narrative \mathcal{N} after removing the observation $\{\neg flying(tweety)\}$ at 5. Figure 1 shows, for some initial time-period, the argument sets $\theta'(T)$ that the agent builds at each time-point T , for one of many possible initial states, with the aim of constructing a model \mathcal{M}' .

For the initial state at time-point 0, the endogenous arguments are largely inapplicable to make predictions — with the exception of the static knowledge and the no-action axiom. Hence, the agent is forced to introduce in $\theta'(0)$ (a minimal set of) exogenous arguments to account (along with the static knowledge) for how state $\mathcal{M}'(0)$ came about.

For the state at time-point 1, all properties hold by persistence or by the no-action axiom, and $\mathcal{M}'(1) = \mathcal{M}'(0)$.

For the state at time-point 2, most properties hold by persistence or by the no-action axiom. To properly account for narrative \mathcal{N}' , the property $shoot(fred)$ at 2 needs to hold in state $\mathcal{M}'(2)$. However, no endogenous argument can account for this, since the action is not *predicted* to occur by the theory of the agent; rather it is simply *observed* to occur. Hence, the action's occurrence needs to be attributed to some exogenous reason. The agent introduces, thus, the exogenous argument $shoot(fred)$ at 2 in $\theta'(2)$.

For the state at time-point 3, most properties hold by persistence or by the no-action axiom. This is not the case for $\neg noise$ and $alive(fred)$ whose persistence is attacked, respectively, by the argument sets $\{C1(fred,3), C1FA(fred,3)\}$ and $\{C2(fred,3), C2FA(fred,3)\}$. The attacks could have been defended had exogenous arguments flipping the preferences in $C1FA(fred,3)$ and $C2FA(fred,3)$ been introduced in $\theta'(3)$. However, that would have resulted in a non-minimal use of exogenous arguments, and by Definition 7, \mathcal{M}' would not have been a model. Hence, instead of exogenously supporting persistence, $\theta'(3)$ includes the endogenous arguments $C1(fred,3)$ and $C2(fred,3)$, as intuitively expected.⁸

For the state at time-point 4, most properties hold by persistence or by the no-action axiom. This is not the case for $\neg flying(tweety)$ that is caused to change. Indeed, the argument $C3(tweety,4)$ is included in $\theta'(4)$, much in the same way that the arguments $C1(fred,3)$ and $C2(fred,3)$ were included in $\theta'(3)$. Persistence cannot be applied to establish *noise* either, since the argument set $\{T1(4), T1FA(4)\}$ offers

violated in favor of satisfying the former (domain-specific) one.

⁸Note that had Fred been able to fly, the static knowledge would have attacked $C2(fred,3)$ through $S5(fred,3)$ and $S5C2(fred,3)$, with no way to defend (without exogenous arguments). This indicates that our domain description needs to be updated so that the $shoot(X)$ action be encoded as a “strong” one, in the same way as $spell(X)$.

an attack that cannot be defended (without exogenous arguments). Hence, the argument $T1(4)$ is included in $\theta'(4)$.

$\theta'(T)$ for subsequent time-points $T \geq 5$ contains arguments coming through persistence and the no-action axiom only, and the state $\mathcal{M}'(T)$ is equal to the state $\mathcal{M}'(4)$.

By Definition 6, \mathcal{M}' is an interpretation of the domain that 1-accepts and agrees with θ' , and also accounts for \mathcal{N}' . By the minimal use of exogenous arguments, and by Definition 7, \mathcal{M}' is a model of the domain for the narrative \mathcal{N}' ; and by Definition 8, \mathcal{M}' is trivially a preferred model.

Unexpected Observations

Let us now consider the original narrative \mathcal{N} . Clearly, \mathcal{M}' is not a model of the domain for \mathcal{N} since, in particular, it does not account for the observation $\{\neg flying(tweety)\}$ at 5. There are many ways to update \mathcal{M}' (and θ') so as to obtain \mathcal{M} (and θ that \mathcal{M} 1-accepts and agrees with) that is a model of the domain for \mathcal{N} . We discuss possible scenarios below:

(i) Tweety is flying at time-point 4, but then lands. This could be modelled by having $\theta(5)$ include, instead of the persistence argument $FA(flying(tweety),5)$, the exogenous argument $\neg flying(tweety)$ at 5. [*Failure of persistence.*]

(ii) The noise did not cause Tweety to fly in the first place. Among many approaches, this could be modelled by having $\theta(4)$ include, instead of the causal argument $C3(tweety,4)$, the persistence argument $FA(\neg flying(tweety),4)$ along with the exogenous exceptional preference $FA(\neg flying(tweety),4) \succ_1 C3(tweety,4)$ stating that persistence was not overridden by the causation of $flying(tweety)$. [*The causal law failed.*]

(iii) Some unexpected occurrence of the *known* shoot action occurred, killing Tweety. This could be modelled by including the exogenous argument $shoot(tweety)$ at 3 in $\theta(3)$, which, with the aid of the static arguments $S5(tweety,4)$ and $S7(tweety,4)$ in $\theta(4)$, and the preferences $S5C3(tweety,4)$ and $S7C3(tweety,4)$, would override / defeat the causation of $flying(tweety)$. [*Exogenous occurrence of a known action (leading to an endogenous qualification of causality).*]

(iv) The noise might not have been caused at all. Among many approaches, this could be modelled by having $\theta(2)$ include, instead of the persistence argument $FA(loaded,2)$, the exogenous argument $\neg loaded$ at 2. [*Failure of persistence.*] Beyond explaining why Tweety is not flying at time-point 5, this explanation also has repercussions on the state of Fred; unlike other scenarios, Fred in this scenario is alive.

(v) Tweety is (or becomes at some point) a penguin. In the simplest case this could be modelled by having $\theta(0)$ contain the exogenous argument $penguin(tweety)$ at 0 instead of its negation. [*Non-deterministic population of the initial state.*]

It is clear, then, that whether Fred is alive after being shot, depends critically on the explanation we are willing to give about why Tweety does not fly after shooting Fred!

Reasoning About Failures

Among the scenarios presented, each mapping θ uses exogenous arguments minimally, and gives rise to a model \mathcal{M} of the domain for the narrative \mathcal{N} . Yet, maybe not all the resulting models should be considered equally preferred.

Typically, for instance, causal failures (ii) are taken to be preferred to persistence / no-action failures (i), (iii), (iv).

time-point	0	1	2	3	4	5
literals (those changing between states shown in bold)	bird(fred) bird(tweety)		bird(fred) bird(tweety)	bird(fred) bird(tweety)	bird(fred) bird(tweety)	bird(fred) bird(tweety)
	turkey(fred) -turkey(tweety)		turkey(fred) -turkey(tweety)	turkey(fred) -turkey(tweety)	turkey(fred) -turkey(tweety)	turkey(fred) -turkey(tweety)
	-penguin(fred) -penguin(tweety)		-penguin(fred) -penguin(tweety)	-penguin(fred) -penguin(tweety)	-penguin(fred) -penguin(tweety)	-penguin(fred) -penguin(tweety)
	-canfly(fred) canfly(tweety)		-canfly(fred) canfly(tweety)	-canfly(fred) canfly(tweety)	-canfly(fred) canfly(tweety)	-canfly(fred) canfly(tweety)
	-flying(fred) -flying(tweety)		-flying(fred) -flying(tweety)	-flying(fred) -flying(tweety)	-flying(fred) -flying(tweety)	-flying(fred) flying(tweety)
	alive(fred) alive(tweety)		alive(fred) alive(tweety)	-alive(fred) alive(tweety)	alive(fred) alive(tweety)	-alive(fred) alive(tweety)
	-noise loaded		-noise loaded	noise loaded	-noise loaded	-noise loaded
	-shoot(fred) -shoot(tweety)		shoot(fred) -shoot(tweety)	-shoot(fred) -shoot(tweety)	-shoot(fred) -shoot(tweety)	-shoot(fred) -shoot(tweety)
	-spell(fred) -spell(tweety)		-spell(fred) -spell(tweety)	-spell(fred) -spell(tweety)	-spell(fred) -spell(tweety)	-spell(fred) -spell(tweety)
arguments (exogenous underlined)	<u>L at 0</u> for some minimal subset of the L above		<u>$shoot(fred)$ at 2</u>	$C1(fred,3)$ $C2(fred,3)$	$C3(tweety,4)$ $T1(4)$	
	$S1(X,0), \dots, S8(X,0)$ for each X <u>$NA(A,0)$</u> for every -A above		$FA(L,2)$ for every L above <u>$NA(A,2)$</u> for every -A above	$FA(L,3)$ for the rest L above <u>$NA(A,3)$</u> for every -A above	$FA(L,4)$ for the rest L above <u>$NA(A,4)$</u> for every -A above	
model	L at 0 for every L above A at 0 for every A above $-A$ at 0 for every -A above		L at 2 for every L above A at 2 for every A above $-A$ at 2 for every -A above	L at 3 for every L above A at 3 for every A above $-A$ at 3 for every -A above	L at 4 for every L above A at 4 for every A above $-A$ at 4 for every -A above	

Figure 1: A model and the corresponding argument sets that support the state transitions.

Among persistence / no-action failures, those occurring due to unexpected occurrences of known actions (iii) might be preferred to those appealing to exogenous actions (i), (iv).

Even among failures of persistence due to exogenous actions, one may impose fluent-specific preferences; e.g., it is more plausible for a bird to land (i) than for a gun to become unloaded (iv) when these changes are not explicitly caused.

It is even conceivable to wish to impose preferences that are time-specific; e.g., exogenous failures are preferred to have occurred earlier (v) than later (i), (ii), (iii), (iv).

Definition 8 can account for such preferences, if we first introduce appropriate \succ_2 comparisons between arguments.

Before discussing the encoding of these \succ_2 preferences, however, let us first consider a slightly different encoding of the domain that includes the spontaneous-change axiom.

$$SC(L,T) \quad : \quad \{\top\} \text{ at } T-1 \rightarrow \{L\} \text{ at } T$$

Typically, spontaneous change is a very weak argument, weaker even than persistence. Indeed, the frame axiom is aimed at encoding that things do not change spontaneously.

$$FASC(L,T) \quad : \quad FA(L,T) \succ_1 SC(-L,T)$$

Under this updated representation, exogenous arguments are not needed to account for the initial state (v). Instead, the endogenous spontaneous-change arguments can be used — since they are not attacked by the persistence arguments $FA(L,0)$, as the conditions $\{L\}$ at -1 of the latter do not hold.

Failures of persistence (i) need not be accounted for by direct exogenous arguments, either. Instead of appealing to the exogenous argument $\neg flying(tweety)$ at 5, one may appeal to the spontaneous-change argument $SC(\neg flying(tweety),5)$, and accompany it by the exogenous exceptional preference $SC(\neg flying(tweety),5) \succ_1 FA(flying(tweety),5)$ to allow spontaneous-change to override persistence in this case.

In effect, these changes in the domain representation aim to *internalize* some of the exogenous arguments that were used to reason in our earlier discussion. Indeed, since our goal is to reason about the use of such exogenous arguments,

it follows that we have some knowledge about them, and that it would be more natural to treat them, to the extent supported by our knowledge, as endogenous arguments. This is exactly what this updated representation achieves.

It is now possible to encode the \succ_2 preferences among the five scenarios. These preferences need to be included both in \mathcal{D} , so as to acknowledge that they are endogenous preferences, but also in θ , so as to allow the associated model to attack other models on the basis of those preferences.

We encode the preference of a given causal law failing over persistence failing, through the following argument, where $C(\cdot, \cdot)$ is the argument encoding the causal law:

$$(FA(L_1, T_1) \succ_1 C(\neg L_1, T_1)) \succ_2 (SC(\neg L_2, T_2) \succ_1 FA(L_2, T_2))$$

Of course, had we wished to restrict the preference to particular literals or time-points, or even restrict it to be applicable in particular settings only, this could have been achieved by making the preference conditional, much like $S4(X,T)$.

By a completely analogous approach, we can encode the rest of the \succ_2 preferences that we have discussed earlier.⁹

Conclusions and Related Work

We illustrated how argumentation offers a natural solution to the knowledge qualification problem. By encoding the reasoning patterns as arguments, the framework is able to override them when and if needed, offering a unified and clear treatment of the endogenous and exogenous qualification.

The proposed framework addresses the challenge of developing an *integrated formalism for reasoning with both default static and default causal knowledge*, a challenge first introduced in [Kakas, Michael, and Miller, 2008a]. Following that initial work and an early attempt at using a model-theoretic approach [Kakas, Michael, and Miller, 2008b], a framework using argumentation as its basis was developed

⁹While the original representation treats models with different initial states (if populated by incomparable exogenous argument sets) as incomparable, the new representation makes such models comparable. Space constraints do not allow further discussion.

[Michael and Kakas, 2009]. Earlier, other works used argumentation also, to address (some of) the frame, ramification, and action qualification problems [Kakas, Miller, and Toni, 1999; Vo and Foo, 2005]. We share with the last three works the use of argumentation for tackling knowledge qualification. We diverge from them in that we do not develop semantics for specific reasoning patterns. Instead, we consider the general knowledge qualification problem, while we also offer a richer solution to the problem of reasoning with unexpected observations, and the associated failure recovery.

Some work has also been done on the use of default reasoning in inferring causal change, and in particular in the case of the action qualification problem [Doherty et al., 1998; Thielscher, 2001]. Other work has investigated the distinction between default and non-default causal rules in the context of a particular action language [Chintabathina, Gelfond, and Watson, 2007]. Following the introduction of the problem of reasoning with default static knowledge in a temporal setting [Kakas, Michael, and Miller, 2008a], there has been some work on developing a logic-based formalism for that problem [Baumann et al., 2010], relying on carefully devised effect axioms to take into account the conclusions of a default static theory. A similar take, but in a modal setting, has also been considered [Lakemeyer and Levesque, 2009].

Future work will be pursued in two main directions. The first direction is to establish formal results for the developed formalism. We wish to: (i) illustrate how certain natural reformulations of the proposed framework are, in fact, equivalent to or subsumed by the present approach (e.g., show how to recover the semantics of existing frameworks that address only some aspects of knowledge qualification); (ii) identify conditions under which the semantics can be further simplified (e.g., when the narrative is online, the narrative can be taken into account during the step-by-step reasoning phase); (iii) identify conditions under which reasoning can be shown to be provably tractable or intractable (e.g., certain syntactic restrictions on the arguments may provably reduce the complexity of computing preferred models); (iv) illustrate that the semantics enjoys typical desirable properties (e.g., it is elaboration tolerant [McCarthy, 1999], and it enjoys the free-will property [Kakas, Michael, and Miller, 2011]). The second direction is to develop an agent that reasons through argumentation, following the pragmatic step-by-step approach suggested herein. We hope that both directions will help us explore and understand further the expressiveness and naturalness of the proposed framework.

References

Baumann, R.; Brewka, G.; Strass, H.; Thielscher, M.; and Zaslowski, V. 2010. State Defaults and Ramifications in the Unifying Action Calculus. In *Proceedings of the 12th International Conference on Principles of Knowledge Representation and Reasoning (KR'10)*, 435–444.

Chintabathina, S.; Gelfond, M.; and Watson, R. 2007. Defeasible Laws, Parallel Actions, and Reasoning about Resources. In *Proceedings of the 8th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense'07)*, 35–40.

Dimopoulos, Y.; Moraitis, P.; and Amgoud, L. 2008. Theoretical and Computational Properties of Preference-Based Argumentation. In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI'08)*, 463–467.

Doherty, P.; Gustafsson, J.; Karlsson, L.; and Kvarnström, J. 1998. TAL: Temporal Action Logics Language Specification and Tutorial. *Electronic Transactions on Artificial Intelligence* 2(3–4):273–306.

Dung, P. M. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games. *Artificial Intelligence* 77(2):321–358.

Kakas, A., and Moraitis, P. 2003. Argumentation-Based Decision Making for Autonomous Agents. In *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'03)*, 883–890.

Kakas, A.; Michael, L.; and Miller, R. 2008a. Fred meets Tweety. In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI'08)*, 747–748.

Kakas, A.; Michael, L.; and Miller, R. 2008b. Fred meets Tweety. In *Proceedings of the 6th International Cognitive Robotics Workshop (CogRob'08)*.

Kakas, A.; Michael, L.; and Miller, R. 2011. Modular-E and the Role of Elaboration Tolerance in Solving the Qualification Problem. *Artificial Intelligence* 175(1):49–78.

Kakas, A.; Miller, R.; and Toni, F. 1999. An Argumentation Framework for Reasoning about Actions and Change. In *Proceedings of the 5th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR'99)*, 78–91.

Lakemeyer, G., and Levesque, H. J. 2009. A Semantical Account of Progression in the Presence of Defaults. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09)*, 842–847.

McCarthy, J. 1999. Elaboration Tolerance. <http://www-formal.stanford.edu/jmc/elaboration/>.

Michael, L., and Kakas, A. 2009. Knowledge Qualification through Argumentation. In *Proceedings of the 10th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR'09)*, 209–222.

Prakken, H., and Sartor, G. 1997. Argument-Based Extended Logic Programming with Defeasible Priorities. *Journal of Applied Non-Classical Logics* 7(1):25–75.

Thielscher, M. 2001. The Qualification Problem: A Solution to the Problem of Anomalous Models. *Artificial Intelligence* 131(1–2):1–37.

Vo, Q. B., and Foo, N. Y. 2005. Reasoning about Action: An Argumentation-Theoretic Approach. *Journal of Artificial Intelligence Research* 24:465–518.