Evidence-Based Belief Revision for Non-Omniscient Agents

Kristina Gogoladze

joint work with Alexandru Baltag

It has long been recognized that inconsistencies may easily occur in people's beliefs in real life. Even if one is rational, one may hold inconsistent beliefs due to receiving conflicting information along with the fact that our limited capacity for information processing (or limited memory) may make it hard to spot the inconsistency. A rational agent would, of course, like to revise his beliefs when he becomes aware of an inconsistency. However, the usual discussion in the Belief Revision literature on solving the contradiction involving old evidence and new evidence assumes that the agent is always aware of this contradiction (because of his logical omniscience).

An important outstanding problem in epistemic and doxastic logics is the problem of logical omniscience, unrealistic assumptions with regard to the reasoning power of the agents. It would be nice, of course, to have perfect reasoners, but even in powerful computers the resources for reasoning are limited. Epistemic logicians usually consider the following features as different issues involving logical omniscience: Knowledge of all logical validities; Monotonicity; Closure under known implication, logical equivalence or conjunction; Introspection. Each of these principles is a feature of idealized perfect reasoners that may not exist in a rational agent in real life.

Assuming that the agent is rational, the reasons he may be non-omniscient are typically limited computational power, time constraints and insufficient memory. These restrictions may also cause the agent to believe some contradictory facts (in this case, he simply may not have noticed the contradiction yet). We are not aware of any previous work that deals with inconsistent beliefs and that has a framework that would allow agents to fix inconsistent beliefs later. There are so-called paraconsistent logics [Priest, 2002] that allow reasoning about inconsistencies, but the underlying philosophy of these logics is that believing a contradiction may be rational and that, in principle, there is no need to resolve logical contradictions. So, if we want to be able to explain why agents can hold inconsistent beliefs, we need to think of something different.

We introduce and investigate a model of belief formation that is closer to real-life reasoning than existing models. In particular, we want to propose a model that enables agents to reason about inconsistent beliefs when they are not aware of the inconsistency due to some limitations by introducing more natural definition of beliefs. Even rational agents may happen to believe irrational things either because they read/were told something or have false evidence from other sources. An agent will never believe an explicit contradiction $\bot$. If he notices such inconsistency, he will have to revise his current beliefs to keep them consistent.

Since one of the main reasons why people hold inconsistent beliefs is limited computational resources, as a possible solution, we, firstly, restrict agents to the usage of only finite amount of sentences at every given period of time. These are going to be (the agent's) *explicit* beliefs — a finite set of syntactically given formulas. *Implicit* beliefs will not have all the restrictions we impose on the explicit beliefs, but agents can reason only with their explicit belief sets. The explicit belief sets are not required to be closed under any of the logical operations, the only restriction will be that they do not contain an *explicit inconsistency* which we denote by $\bot$. Then we go one level up and start with *explicit evidence pieces* instead of explicit beliefs by borrowing some ideas from van Benthem and Pacuit's work [2011] on evidence-based beliefs. The explicit beliefs of an agent will then be computed using his explicit evidence pieces. This

"computation" is defined in such a way that it does not allow an explicit inconsistency in the agent's explicit belief set even when his evidence set does contain $\bot$.

Interestingly, our proposed models of explicit beliefs naturally validate axiom schemes that correspond to nice properties of knowledge and belief that one may want to have.

Definition 1 (Explicit Evidence Language)**.** Let At be a set of atomic propositions. Formulas ϕ of language $\mathcal{L}$ are given by

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid B^e\phi \mid B^i\phi \mid K^e\phi \mid K^i\phi$$

with $p \in \mathsf{At}$. ◁

We use the B^e and B^i modalities for explicit and implicit belief respectively, and, similarly, K^e and K^i for explicit and implicit knowledge.

Definition 2 (Semantic Model of Explicit Evidence)**.** An *explicit evidence model* (EE-model) is a tuple $\mathfrak{M} = (W, W_0, \mathcal{E}_s, \mathcal{E}_h, V)$ where

W is a set of possible worlds.

$W_0 \subseteq W$ is a set of worlds that represents the agent's background beliefs or "biases".

$\mathcal{E}_s \subseteq \mathscr{P}(\mathcal{L})$ is a set of formulas that represent agent's (soft) evidence pieces.

$\mathcal{E}_h \subseteq \mathcal{E}_s$ is a set of hard evidence pieces.

$V : \mathsf{At} \to \mathscr{P}(W)$ is a valuation function.

The following condition is imposed on the models:

$$\top \in \mathcal{E}_h(w)$$

◁

The above condition means that the agent has some knowledge to start with. Note that here the soft evidence set may contain $\bot$ (but the explicit beliefs will not). We will use $\bot$ to "mark" a contradiction, but this is only a convention because we did not want to restrict ourselves — it could have been any formula.

The idea is that we can now think of the explicit knowledge set also as of an evidence set — it is hard evidence set $\mathcal{E}_h$ that is infallibly true, whereas $\mathcal{E}_s$ is a soft evidence set: an agent is not absolutely certain about those evidence pieces and they may even be inconsistent with each other.

Definition 3 (Quasi-consistency)**.** Let U be a set of formulas. We say that U is *quasi-consistent* if $\bot \notin U$. ◁

Definition 4 (Closed Evidence)**.** A set $F \subseteq \mathcal{E}_s$ of (soft) evidence pieces is said to be *closed* if, and only if, it includes all the hard evidence (i.e. $\mathcal{E}_h \subseteq F$) and it is closed under Modus Ponens within $\mathcal{E}_s$ (i.e. if ϕ and $\phi \to \psi$ belong to F and ψ belongs to $\mathcal{E}_s$, then ψ belongs to F). ◁

Definition 5 (Q-max Evidence)**.** A set $F \subseteq \mathcal{E}_s$ of (soft) evidence pieces is said to be *maximal closed quasi-consistent set* (or *q-max*, for short) if it is (1) closed (in the above sense), (2) quasi-consistent, and (3) maximal with respect to properties (1) and (2) (i.e. for every other closed quasi-consistent set F', if $F \subseteq F' \subseteq \mathcal{E}_s$, then $F' = F$). ◁

Since we do not have explicit beliefs in the model, we have to define them. Soft evidences are on the more abstract level than beliefs, the evidence pieces play a role of the derivations an agent made so far, and beliefs are encoded there. We say that the agent *explicitly believes* a formula at some world if, and only if, that formula belongs to the intersection of all maximal closed quasi-consistent sets:

$$\mathcal{B} \coloneqq \bigcap \{F : F \text{ is q-max}\}$$

Let us use this abbreviation for the explicit belief set from now on.

The choice of such definition naturally arises from our line of research — since we assume the agent has the fast "working" memory where he can easily compute even exponential things. According to this definition, the agent stays safe and cautious and sticks with what is included to every maximal closed quasi-consistent set of evidence pieces. This can be seen as the appropriate syntactic counterpart of the van Benthem and Pacuit's definition of Maximal Consistent Evidence [2011]. Required $\bot \notin \mathcal{B}$ will hold automatically by construction.

In our models, implicit belief is a defined notion: it is defined as a closure of agent's explicit beliefs together with his prior background biases. From this it follows that implicit beliefs may be inconsistent (in the usual sense). We think, that defining implicit beliefs via explicit ones is more natural than treating them as an independent notion, and it makes perfect sense that the implicit beliefs of an agent may happen to be inconsistent at some point in time. Of course, these beliefs can become consistent if the agent manages to resolve the inconsistencies in his explicit beliefs.

Working with the assumption that the agent remains rational, and that he does not find it rational to believe in explicit inconsistencies, we provided a model that in a sense corrects the explicit inconsistencies itself.

Our proposed solution addresses this problem by providing a basis for agent's beliefs — syntactic pieces of evidence that an agent uses to justify his beliefs. Then, the explicit beliefs of an agent are computed using his explicit evidence set. The explicit belief set is purely syntactic as well, which allows an agent to hold any kind of sentences without identifying them with an inconsistency, unless it is indeed an explicit inconsistency.

Since we allow our agents to operate only with (finite) syntactic explicit information, our proposed explicit evidence models happen to resolve all the omniscience problems that epistemic logicians are usually concerned with. Closure properties need not hold at all for the explicit sets of knowledge and belief, as well as explicit introspection.

We prove the following theorem.

Theorem. *The logic is completely axiomatized by the following system of axioms and rules:*

$S5$ *axioms and rules for* K^i	(1)	$B^e\phi \to B^i\phi$	(2)
$K45$ *axioms and rules for* B^i	(3)	$K^e\phi \to K^iK^e\phi$	(4)
$\neg B^e\bot$	(5)	$\neg K^e\phi \to K^i\neg K^e\phi$	(6)
$K^e\top$	(7)	$B^e\phi \to K^iB^e\phi$	(8)
$K^i\phi \to B^i\phi$	(9)	$\neg B^e\phi \to K^i\neg B^e\phi$	(10)
$K^e\phi \to B^e\phi$	(11)	$B^i\phi \to K^iB^i\phi$	(12)
$K^e\phi \to K^i\phi$	(13)	$\neg B^i\phi \to K^i\neg B^i\phi$	(14)

Here, reflexivity of knowledge represents its *veracity*, transitivity — *positive introspection*, and Euclideanness — *negative introspection*; as for belief, transitivity means positive introspection, Euclideanness — negative introspection, and seriality would mean *consistency*; there

is also *strong introspection* of belief (axioms 8,10,12 and 14), and the obvious requirement that knowledge should imply belief (axioms 9 and 11).

It is worthwhile to mention that neither B^e nor B^i satisfy the standard $KD45$ axioms for belief. Implicit beliefs do not satisfy the seriality axiom, as explained above, whereas explicit beliefs do not satisfy the K-axiom (The axiom K means that the set of formulas that the agent knows is deductively closed. It would imply logical omniscience of the agent.). Interestingly, B^e does satisfy the D-axiom in the sense that $\neg B^e \bot$ (but not $B^e\phi \to \neg B^e \neg\phi$). Consequently, one could argue that the standard notion of belief is a mixture of these two.

One of our main intentions in this work was to allow agents to resolve the inconsistency once they become aware of it. To model this, we have to express some *actions* that describe how the agent becomes aware of new information. We focus on some of the possible evidence dynamics which are also called *updates*. The updates are informational actions that *change* the original model. Some of these changes may remove the possible worlds of the agent, another — will just modify the explicit information of the agent. First of all, one could look at the usual DEL [van Ditmarsch *et al.*, 2007] updates. For example, the operation of update models the situation when the agent receives a piece of evidence from an infallible source. Another, more natural for our models, scenario is when the agent learns new pieces of evidence. They may be consistent or inconsistent with the previously learned information. It can be even explicit inconsistency $\bot$. There are two possibilities: either the agent adds ϕ to his explicit knowledge, or he just accepts ϕ as a piece of evidence. We would like to model belief revision of realistic agents, and realistic agents cannot hold all the information they learn forever. In real life, agents do forget some things from time to time. It, therefore, makes perfect sense to consider evidence removal operation as well. With these dynamic operations, the agent can become aware of the inconsistency and is able to fix it (this happens automatically). This means that both explicit and implicit beliefs of the agent can become consistent (if they were not).

We have given an axiomatization for the (static) logic of explicit evidence which is complete. Next, we presented the dynamic actions that describe change of models due to modifications in the evidence pieces. We saw various examples that illustrate how our models work. In the key part of this work, we showed how the problem of inconsistent belief revision is solved with the help of our models. Lastly, we discussed some possible extensions of the language of explicit evidence in order to provide sound and complete system for the extended dynamic language.

References

[van Benthem and Pacuit, 2011] J. van Benthem and E. Pacuit. Dynamic logics of evidence-based beliefs. *Studia Logica*, 99:61–92, 2011.

[van Ditmarsch *et al.*, 2007] H. van Ditmarsch, W. van Der Hoek, and B. Kooi. *Dynamic epistemic logic*. Springer, 2007.

[Fagin and Halpern, 1987] R. Fagin and J.Y. Halpern. Belief, awareness, and limited reasoning. *Artificial intelligence*, 34(1):39–76, 1987.

[Priest, 2002] G. Priest. *Handbook of Philosophical Logic*, volume 6, chapter Paraconsistent logic, pages 287–393. Kluwer Academic Publishers, 2002.

[Velázquez-Quesada, 2014] F. R. Velázquez-Quesada. Dynamic epistemic logic for implicit and explicit beliefs. *Journal of Logic, Language and Information*, 23(2):107–140, 2014.