Comments on the PhD thesis "Trustworthiness and Expertise: Social Choice and Logic-based Perspectives" by Joseph Singleton (Univ. Cardiff)

Andreas Herzig, CNRS, Univ. Toulouse

The thesis is situated in the domains of artificial intelligence and multi-agent systems; more precisely, it is about truth discovery in situations where sources with varying degrees of expertise provide possibly conflicting claims, where the latter are pieces of information about the properties of some objects. The task then is to evaluate the trustworthiness of the sources and the believability of their claims.

Joseph Singleton's PhD thesis takes a formal logical approach: he formulates axioms that any rational truth discovery method should satisfy and he provides a logical analysis of two central concepts: expertise of the sources and soundness of their claims. His logical models of expertise and soundness are inspired by epistemic logics and his account of their evolution is inspired by dynamic epistemic logics and belief revision theory.

Overview of the Contributions

The thesis is organised in 5 chapters, plus an introduction (Chapter 1) and a conclusion (Chapter 2). It contains three main contributions:

- Rationality principles for truth discovery (Chapter 2) as well as its relationship with the graph-theoretic notion of (bipartite) tournaments (Chapter 3);
- A non-standard semantics of a modal logic of expertise and soundness of information, its axiomatisation, dynamic extension, and its relationship with standard epistemic logic (Chapter 4);
- A generalisation of existing approaches to the revision of beliefs to the case where different sources provide possibly contradicting information, taking varying expertise of the sources and trust in these sources into account (Chapter 5); plus the application of that framework to truth tracking (Chapter 6).

All these contributions are important and substantial. All results are proved rigorously, and I was not able to spot any error or omission.

In the rest of the report I will go through the chapters one-by-one and provide some comments and criticisms. I would like to say right away that all my criticisms have to be taken as minor.

Chapter 2

Is about an axiomatic analysis of truth discovery in the style of social choice theory. It is based on truth discovery networks $N = \langle S, O, D, R\rangle$ where S are the sources, O are the objects, $\{D_o\}_{o\in O}$ is the domain of objects ('values'), with $V = \bigcup_{o\in O} D_o$, and $R \subseteq S \times O \times V$ are the reports.

Comments and questions:

1. There is a constraint that R is functional for the value argument: why not formulate the 'report' relation R as a partial function $S \times D \longrightarrow V$?

2. The voting operator framework T_N^{vote} on Page 21 allows for untrustworthy sources s whose weights $w_N(s)$ are smaller than 0 and therefore decrease the credibility of a claim. One should make a difference here between 'ignorers' and 'liars': the former should not influence the credibility of a claim while the latters should *decrease* it. This is only acknowledged on page 56 where limitations are mentioned: "simplifying assumptions we make regarding source attitudes; e.g. that they do not collude or attempt to manipulate".

   This is also related to the discussion on Page 31 where the 'responsiveness' axiom **Fresh-pos-resp** is introduced and motivated. It is said that "we have no reason to believe [that fresh] sources are untrustworthy, and they should therefore have a positive effect when making a claim". But actually we have no reason either to believe that fresh sources are trustworthy! It follows that even the weakening of the responsiveness axiom as motivated here is too optimistic. Suppose for example that the newcomer votes for the claim with lowest score: this could indicate that she is a liar, and it would be more prudential to wait until her status is better known.

   The same criticism applies to axiom **Source-pos-resp**.

Chapter 3

Relates truth discovery to the graph-theoretical notion of bipartite tournaments. There, the points are partitioned into two sets such that there is no edge between members of the same set. To give an example, the problem is to establish a national ranking of football teams which only play international matches against teams from some other country and therefore have to be compared indirectly. Then one country corresponds to the sources and the other one to the claims. The ranking is clear when the bipartite tournament is a chain graph, and the method consists in finding the minimum number of edge changes so that a chain graph is obtained. This corresponds to finding minimal Hamming distances between graphs.

Comments and questions: I have no particular comments on this chapter; except the somewhat speculative:

1. Is there any relation between bipartite tournaments and stable marriage problems?

Chapter 4

Provides a kind of epistemic logic in terms of a non-standard possible world semantics: expertise models. There are three modal operators $\forall$, $\mathbf{E}_i$ and $\mathbf{S}_i$, where $\forall\varphi$ is a universal modal operator (written $\mathbf{A}$ in the thesis); $\mathbf{E}_i\varphi$ means that source i is an expert on φ; and $\mathbf{S}_i\varphi$ means that source i is sound about φ.

The expertise models have sets of sets of possible worlds P_i associated to each source i. $\mathbf{E}_i\varphi$ is true at possible world x if the extension of φ, $||\varphi||_M$, is in P_i; $\mathbf{S}_i\varphi$ is true at possible world x if for every $A \in P_i$ containing the extension of φ, $x \in A$. For example, if $P_i = \{||p||_M\}$ then $\mathbf{E}_i p$ is true at every x, and $\mathbf{S}_i(p \wedge q)$ is true at every x where p is true.

Various restrictions on the set P are considered, in particular closure of P under intersection and union. They make that $\mathbf{E}_i\varphi$ not only implies, but is even equivalent to $\forall(\mathbf{S}_i\varphi \to \varphi)$.

Comments and questions:

1. The term 'soundness' is perhaps not the best reading for the modal operator S_i; 'consistency (with the source's knowledge)' might be more appropriate, in particular given the identity with the epistemic diamond operator that is established in Chapter 4.3 (under the hypothesis of closure of P under intersection and union).

   Wouldn't "φ is consistent with i's expertise" be better?

2. The intuitions that is given about the expertise operator E_i is that "a source has expertise on φ if they are able to correctly refute φ in any situation where it is false" (page 108). Similarly, the interpretation of $A \in P_i$ that is given on Page 111 is that "whenever the "actual" state is outside A, the source knows so". This would mean that the formula $E_i\top$ should be valid, but it is not.

3. As an aside, I wondered whether the logic could shed a new light on the knowability paradox. Formally, the knowability thesis is expressed by the formula $\varphi \to \neg\forall\mathbf{S}_i\neg\varphi$, and when φ is instantiated by $p \wedge \mathbf{S}_i\neg p$ then one obtains

   $$p \wedge \mathbf{S}_i\neg p \to \neg\forall\mathbf{S}_i\neg(p \wedge \mathbf{S}_i\neg p)$$

   whose right-hand-side is unsatisfiable due to the T_S-axiom for knowledge.

   Is it in line with the knowability paradox—or perhaps even equivalent to it—to say that an agent cannot have expertise about her own ignorance?

4. The only introspection principle that is discussed is $\mathbf{S}_i\mathbf{S}_i\varphi \leftrightarrow \mathbf{S}_i\varphi$. What about other nestings such as $\mathbf{E}_i\mathbf{E}_i\varphi$ and $\mathbf{S}_i\mathbf{S}_i\varphi$? It is only said on page 151 (footnote 1) that when P is closed under intersections and unions then $\mathbf{E}_i\mathbf{E}_i\varphi$ and $\mathbf{E}_i\mathbf{S}_i\varphi$ are valid and $\mathbf{S}_i\mathbf{E}_i\varphi$ is equivalent to $\mathbf{E}_i\varphi$.

   For example, $\mathbf{E}_i\mathbf{E}_i\varphi$ should be equivalent to

   $$(\mathbf{E}_i\top \wedge \mathbf{E}_i\varphi) \vee (\mathbf{E}_i\bot \wedge \neg\mathbf{E}_i\varphi).$$

   Is that correct?

5. Concerning the axiomatisation of Table 4.1 (Page 122), why not replace the axiom (W_E):

$$\forall(\varphi \to \psi) \to (\mathbf{S}_i\varphi \wedge \mathbf{E}_i\psi \to \mathbf{S}_i\psi)$$

by $\mathbf{S}_i\varphi \wedge \mathbf{E}_i\varphi \to \varphi$? That formula is a consequence of W_E and $\forall(\varphi \to \varphi)$ and is simpler. And indeed, one can derive W_E from it:

(a) $(\forall(\varphi \to \psi) \wedge \mathbf{S}_i\varphi) \to \mathbf{S}_i\psi$ (W_S)

(b) $(\mathbf{S}_i\psi \wedge \mathbf{E}_i\psi) \to \psi$ (new axiom)

(c) $(\forall(\varphi \to \psi) \wedge \mathbf{S}_i\varphi \wedge \mathbf{E}_i\psi) \to \psi$ (from a and b by CPC)

6. On page 117 it is said that the negative introspection axiom **5** "is a rather idealised property of knowledge" and that "it is certainly reasonable to expect that **5** may fail for agents who are not perfectly rational (e.g. humans)".

   Failure of negative introspection is actually a more serious problem than that: it is also undesirable for rational agents, as soon as they get into situations where they wrongly believe to know something. This is described formally by

$$\varphi \wedge \neg\mathbf{K}_i\varphi \wedge \mathbf{B}_i\mathbf{K}_i\neg\varphi = \varphi \wedge \mathbf{S}_i\neg\varphi \wedge \mathbf{B}_i\neg\mathbf{S}_i\varphi,$$

   where $\mathbf{B}_i$ is a modal operator of belief. Then negative introspection tells us that the 2nd conjunct $\neg\mathbf{K}_i\varphi$ implies $\mathbf{K}_i\neg\mathbf{K}_i\varphi = \neg\mathbf{S}_i\neg\mathbf{S}_i\neg\varphi$, which under the 'knowledge implies belief' principle is inconsistent with the 3rd conjunct $\mathbf{B}_i\mathbf{K}_i\neg\varphi = \mathbf{B}_i\neg\mathbf{S}_i\varphi$.

7. In Section 4.3 (specifically on page 121) different properties of knowledge are discussed in terms of closure properties, but not the confluence property of the epistemic logic S4.2.

   Is there a closure property corresponding to S4.2?

8. A general question about the proofs in Chapter 4: Why not put the axiomatics to work and prove theoremhood, instead of proving validity? For example, it seems that the right-to-left direction of Proposition 4.5.2 on page 137 is a direct consequence of the induction axiom for common knowledge

$$\mathbf{K}_J^{\mathsf{com}}(\varphi \to \mathbf{K}_J^{\mathsf{sh}}\varphi) \to (\varphi \to \mathbf{K}_J^{\mathsf{com}}\varphi)$$

   This also applies to Chapter 5, e.g. the proof of Proposition 5.1.2.

9. Proposition 4.6.2 on page 144 lists reduction axioms for the modal operator of sound announcements $[?\varphi]$. It is not said whether these axioms form a *complete* set of reduction axioms. This is actually not the case: there are no reduction axioms for the cases $[?\varphi][?\psi]\chi$ and $[?\varphi][+\psi]\chi$; similarly for the reduction axioms for the modal operator of expertise increase on page 142.

It therefore fails to hold that the axiomatics for L together with the reduction axioms of Proposition 4.6.2 is a complete axiomatisation.

There are two ways out: the first one is to add reduction axioms for the above cases; the second one is to add so-called rules of equivalence for the modal operator of sound announcements

$$\frac{\psi \leftrightarrow \psi'}{[?\varphi]\psi \leftrightarrow [?\varphi]\psi'}$$

and similarly for the modal operator of expertise increase. These inference rules allow then to derive the rule of replacement of equivalents

$$\frac{\psi \leftrightarrow \psi'}{\varphi(\psi) \leftrightarrow \varphi(\psi')},$$

which in turn allows to apply the above reduction axioms inside a formula. Such 'inside out' reduction does not require reduction axioms for combinations of dynamic modalities such as $[?\varphi][?\psi]$.

Chapter 5

Puts to work the logic of Chapter 4 in order to form beliefs given a sequence σ of reports of the form $\langle i, c, \varphi\rangle$ where i is a source, c is a case, and φ is a boolean formula. From contradicting reports from two different sources about the same φ one can then deduce the conclusion that one of the sources fails to be an expert on φ. It is supposed that there is a special source $*$ who is a 'super-expert' with expertise on any φ; this device makes that one can not only deduce non-expertise, but also expertise of another source about φ.

It is supposed that all reports $\langle i, c, \varphi\rangle$ are about a fixed point in time, which means that we are closer to belief revision than to belief update. As Delgrande et al. (2006) argued, in that case the AGM success postulate's 'priority to the input' does not make sense; the framework here adopts that position, which makes that the sequences σ are actually multisets.

Two kinds of conclusions are drawn from a sequence of reports σ: knowledge K^σ and belief B^σ. Epistemic conclusions are basically obtained by translating all φ from the reports $\langle i, c, \varphi\rangle \in \sigma$ into formulas $\mathbf{S}_i\varphi$ of the logic of Chapter 4, and a set of postulates is designed which completely characterises this unique way of obtaining epistemic conclusions. There is more freedom when it comes to computing doxastic conclusions, and several postulates are introduced and discussed. It is shown that one can in particular get belief revision operators tht satisfy the AGM postulates by associating inputs with the priviledged source $*$ (Section 5.4, page 176).

Comments and questions:

1. The logical language of Section 5 restricts that of Section 4 to non-nested formulas. The equal status of sources and objects in Section 2 and 3 contrasts with the language here: operators of expertise and soundness have a source as index, while objects, alias cases, are not first-class citizens: they only appear in the meta-language, as indexes of belief sets $\{\Gamma_c\}_{c\in\mathcal{C}}$.

2. One page 151, it sounds a bit odd to call v_c "the 'true' valuation at case $c \in \mathcal{C}$". Wouldn't it be better to say that v_c are the properties of case c?

3. The semantics in Section 5.1 (page 151) is reminiscent of Pawlak's rough sets, where the semantics has partitions Π as well, and where the unique cell $\Pi[U]$ of Π containing the set of cases U is called the upper approximation of U.[1]

4. Why not present reports as multisets (and even sets) instead of sequences, given that there is the Rearrangement Postulate (and later the Duplicate-removal Postulate)?

5. Again a general question about the proofs in Chapter 5: Why not put the axiomatics to work and prove theoremhood, instead of proving validity? I suppose that the proof of Propostion 5.1.2 (page 154) could be written more concisely.

6. A detail: Page 156 defines consistency as having at least one model; "satisfiability" would be the appropriate term.

7. As noted on page 159, Delgrande et al. (2006) pointed out the tension between revision being about a single time point and the priority to the input postulate. Could it be claimed that the present expertise-based truth-tracking (about a fixed point in time) is the appropriate correction to AGM?

8. The 2018 paper by Delgrande, Peppas and Woltran about generalised AGM belief revision is cited on page 165, and a footnote there notes that "while the result is similar, our framework is not an instance of theirs". The comparison might be detailed a bit more.

Chapter 6

Concludes the thesis by getting back to the problem of truth tracking. As the last paragraph of Chapter 5 says, its postulate-based approach only addresses coherence of the reports among themselves: it does not tell us anything about whether the conclusions that are obtained in this way are indeed true or not.

Previous work of Baltag et al. (2016, 2019) is extended to the setting of Chapter 4 in terms of expertise and sound reports. Just as in Chapter 5, nested

[1] Pulak Samanta, Mihir Kumar Chakraborty: Interface of Rough Set Systems and Modal Logics: A Survey. Trans. Rough Sets 19: 114-137 (2015)

formulas are excluded. Streams of such reports (as customary from formal learning theory) are studied, and a postulate-based analysis similar to that of Chapter 5 is undertaken.

Comments and questions:

1. It seems that all examples in the thesis have reports where the propositional formula that the source provides is a literal, that is, either an atom or its negation. One might consider to make that restriction systematically: would the formal framework get simpler? In particular, would it improve the complexity of solving some decision problems?

2. Just as Chapter 3, the language disallows higher-order knowledge and belief. A puzzle where higher-order knowledge and the distinction between 'ignorers' and 'liers' are crucial is the puzzle "Two doors with two guards". The puzzle starts as follows:

   > You are a prisoner in a room with two doors D_1 and D_2 and two guards. One of the doors will guide you to freedom and behind the other is a hangman–you don't know which is which. One of the guards always tells the truth and the other always lies, but you don't know which one.

   The goal of the original puzzle is to learn the right door by asking a single question to one of the guards. Here one might consider instead the variant where we have two successive reports of each guard both claiming that it is door D_1 leading to freedom. Then one should be able to conclude that it is rather the other door D_2 that leads to freedom.

   This cannot be expressed in the formalism of Chapter 5, but at least in principle it could be expressed in that of Chapter 2.

   But first of all, is this an interesting truth tracking problem? Are there other examples where higher-order expertise and knowledge plays a role in truth tracking?

Concluding Remarks

Format and presentation of the thesis are excellent. The style is appropriate. It is very well-written and is readable by anybody with a background in formal models for multiagent systems; more precisely, in social choice theory, modal logic, and belief revision. Explanations and examples are provided as needed, while avoiding redundancies and overly long developments.

Each of the chapters is backed by one or more publications. Most of them were published in prestigious places such as the AAMAS, IJCAI and KR conferences, which confirms the high quality of Joseph Singleton's contributions.