Engineering Social Order in Multi-Agent Systems

Stephen Cranefield

School of Computing, University of Otago
Dunedin, New Zealand

EMAS@AAMAS 2024
7 May 2024

This work was supported by the Marsden Fund Council from New Zealand Government funding, managed by Royal Society Te Apārangi.

Engineering social order via peer-to-peer interactions

My work has focused on what *agents* themeselves (might) need to coordinate/cooperate, rather than creating centralised services or hierarchical social structures. A few examples:

- Individual modelling and monitoring of social expectations, e.g. a monitor service for Jason agents and application to the Second Life virtual world.
- Learning existing norms from observation of interactions in a society: data mining and Bayesian approaches.
- Choosing agent plans to maximise a human partner's value fulfilment.
- Extending BDI agents to follow (predefined) social practices.
- Peer to peer proposal and execution of group plans, supported by decentralised middleware.
- And now . . . individual agent reasoning about common knowledge.

Examples of common knowledge (within varying groups)

Examples of common knowledge (within varying groups)

- There are four seasons in a year.

Examples of common knowledge (within varying groups)

- There are four seasons in a year.
- After lightening we will hear thunder.

Examples of common knowledge (within varying groups)

- There are four seasons in a year.
- After lightening we will hear thunder.
- Christopher Luxon is the prime minister of New Zealand.

Examples of common knowledge (within varying groups)

- There are four seasons in a year.
- After lightening we will hear thunder.
- Christopher Luxon is the prime minister of New Zealand.
- Taylor Swift and Joe Alwyn have broken up.

Examples of common knowledge (within varying groups)

- There are four seasons in a year.
- After lightening we will hear thunder.
- Christopher Luxon is the prime minister of New Zealand.
- Taylor Swift and Joe Alwyn have broken up.
- 

Examples of common knowledge (within varying groups)

- There are four seasons in a year.
- After lightening we will hear thunder.
- Christopher Luxon is the prime minister of New Zealand.
- Taylor Swift and Joe Alwyn have broken up.
- 

- The band rocks.

Concert crowd image: `https://pxhere.com/fr/photo/727825` (CC0)

Informal definition of common knowledge

Proposition φ is common knowledge if:

- everyone knows φ
- everyone knows that everyone knows φ
- everyone knows that everyone knows that everyone knows φ
- ...

The coordinated attack problem (Fagin et al.)

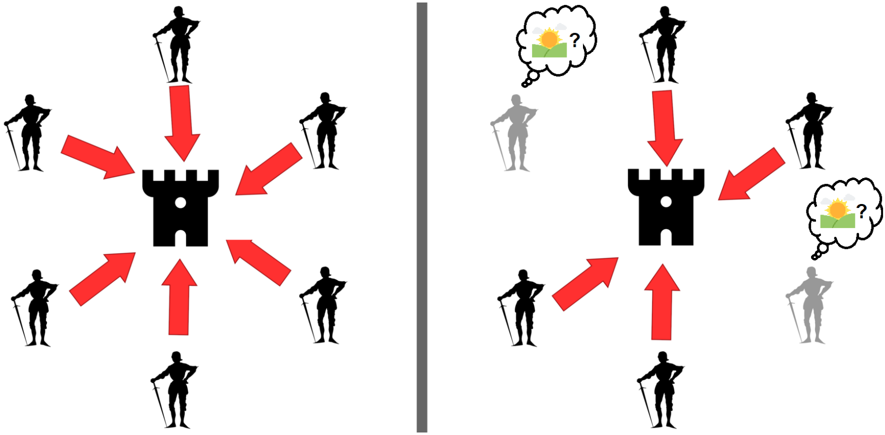
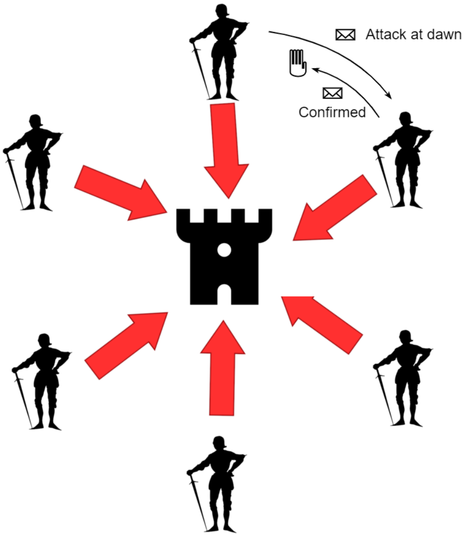


Image derived from Byzantine Generals.png by Lord Belbury. Licence CC BY-SA 4.0 `https://en.wikipedia.org/wiki/Byzantine_fault#/media/File:Byzantine_Generals.png`.
Dawn icon by Freepik, `https://www.flaticon.com/free-icons/dawn`.

The coordinated attack problem (Fagin et al.)

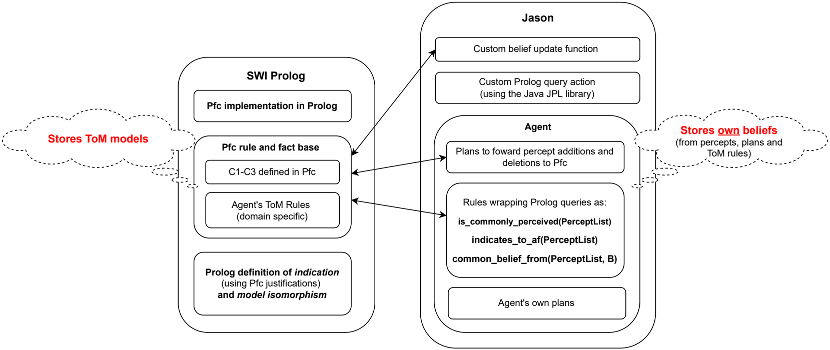


Common knowledge cannot be achieved via asynchronous communication

Why do we need common knowledge?

- Basic assumption of game theory: the payoff structure and the rationality of all players are common knowledge
- Conventions (Lewis). In instances of a coordination problem S, it is *common knowledge* that:
    - There is some regularity of behaviour R that everyone conforms to.
    - Everyone expects everyone else to conform to R.
    - Everyone prefers to conform to R on condition that the others do since R is a coordination equilibrium in S.
- Definition of "We-mode" thinking in a group (e.g. Tuomela 2007)
- ...

Where I am heading: the agent engineering outcome



(Some) theories of common knowledge: 1) Extension of epistemic logic (Fagin et al.)

$K_i\varphi$ means "Agent i knows φ"

$$E_G\varphi := \bigwedge_{i \in G} K_i\varphi\text{: Everyone in group } G \text{ knows } \varphi$$

$$C_G\varphi := \bigwedge_{i=0}^{\infty} E^i\varphi\text{: It is common knowledge in } G \text{ that } \varphi$$

where $E^n_G\varphi := E_G E^{n-1}_G$ and $E^0_G\varphi := \varphi$

To avoid an infinite conjunction, $C_G\varphi$ can be defined using the Fixed-Point Axiom:

$$C_G\varphi \leftrightarrow E_G(\varphi \wedge C_G\varphi)$$

and induction rule[1]:

$$\text{If } \varphi \rightarrow E_G(\psi \wedge \varphi) \text{ then infer } \varphi \rightarrow C_G\psi$$

[1] "The antecedent gives us the essential ingredient for proving, by induction on k, that $\varphi \rightarrow E^k_G(\psi \wedge \varphi)$ is valid for all k" (Fagin et al.)

Problems (?) with the epistemic logic approach

- Artemov (2004):
  - "This kind of deductive system does not behave well proof-theoretically. This practically rules out automated proof search and severely limits the usage of formal methods in analyzing knowledge."
  - "...this [is] the most liberal version of knowledge operator satisfying the Fixed Point Axiom, without imposing any conditions on the way this knowledge is attained. ...there might be nonconstructive versions of the common knowledge appearing by chance or for some unknown reasons or without any particular reasons at all."
- We can (e.g.) combine a logic of common knowledge with *public announcement logic*, with an inference rule that infers common knowledge from a public announcement[2]
  - But how do we decide *what* counts as a public announcement?
  - Do we need add-on logics for other ways of creating common knowledge?
- Why don't see practical agent systems using it.

[2] https://plato.stanford.edu/entries/dynamic-epistemic/#CommKnow

(Some) theories of common knowledge: 2) David Lewis (1969)

- Lewis focuses on situations when a certain *state of affairs A* "indicates" that a proposition *P* holds.
  - Example (Lewis):

    You said you will return tomorrow to continue our meeting
    indicates that
    you will return.
  - Example (Cubitt & Sugden):

    The room we are in is lit by a flash of lightening
    indicates that
    within a few seconds, there will be the noise of thunder.
- Lewis defines (informally) three *sufficient* conditions for the indicator *A* to be a *basis for common knowledge of P*.

(Some) theories of common knowledge: 2) David Lewis (1969)

Proposition P is *common knowledge* if and only if there is some state of affairs A that holds and:

$$\text{Everyone has reason to believe that } A \text{ holds.} \tag{C1}$$

$$A \text{ indicates to everyone that everyone has reason to believe that } A \text{ holds.} \tag{C2}$$

$$A \text{ indicates to everyone that } P \text{ holds.} \tag{C3}$$

Plus “suitable ancillary premises regarding our rationality, inductive standards, and background information”

(Some) theories of common knowledge: 2) David Lewis (1969)

Proposition P is *common knowledge* if and only if there is some state of affairs A that holds and:

Everyone has reason to believe that A holds. (C1)

A indicates[1] to everyone that everyone has reason to believe that A holds. (C2)

A indicates to everyone that P holds. (C3)

Plus "suitable ancillary premises regarding our rationality, inductive standards, and background information"

[1] A indicates φ := If someone has reason to believe that A holds, they *thereby* have reason to believe φ (discussed later).

(Some) theories of common knowledge: 2) David Lewis (1969)

Proposition P is *common knowledge* if and only if there is some state of affairs A that holds and:

Everyone has reason to believe that A holds. (C1)

☞ Cubitt & Sugden: "A is self-revealing"

A indicates to everyone that everyone has reason to believe that A holds. (C2)

☞ C&S: "A is public"

A indicates to everyone that P holds. (C3)

☞ Me: "A is objective"

Plus "suitable ancillary premises regarding our rationality, inductive standards, and background information"

(Some) theories of common knowledge: 2) David Lewis (1969)

- Informal proof: Given some A and P such that C1, C2 and C3 hold, there exists an infinite chain of reasoning that creates all levels of nested reasons to believe:

  i has reason to believe that j has reason to believe that k has reason to believe ... that P.
- The proof doesn't depend on the content of A and P—just the properties C1, C2 and C3.
- The proof can be recast using mathematical induction There is no need to perform an infinite chain of reasoning.

Lewis's informal analysis

II | Convention Refined

1. Common Knowledge

Agreement, salience, or precedent, we have seen, can solve a coordination problem by producing a system of concordant first- and higher-order mutual expectations. We need only imagine cases to convince ourselves that higher-order expectations *would* be produced. But how? What premises have we to justify us in concluding that others have certain expectations, that others expect others to have certain expectations, and so on? And how is the process cut off—as it surely is—so that it produces only expectations of the first few orders?

Take a simple case of coordination by agreement. Suppose the following state of affairs—call it *A*—holds: you and I have met, we have been talking together, you must leave before our business is done; so you say you will return to the same place tomorrow. Imagine the case. Clearly, I will expect you to return. You will expect me to expect you to return. I will expect you to expect me to expect you to return. Perhaps there will be one or two orders more.

What is it about *A* that explains the generation of these higher-order expectations? I suggest the reason is that *A* meets these three conditions:

(1) You and I have reason to believe that *A* holds.
(2) *A* indicates to both of us that you and I have reason to believe that *A* holds.
(3) *A* indicates to both of us that you will return.

What is indicating? Let us say that *A indicates* to someone *x* that

____ if and only if, if *x* had reason to believe that *A* held, *x* would thereby have reason to believe that ____. What *A* indicates to *x* will depend, therefore, on *x*'s inductive standards and background information.

The three main premises (1), (2), (3), together with suitable ancillary premises regarding our rationality, inductive standards, and background information, suffice to justify my higher-order expectations. Let us see how my reasoning would work.

Consider that if *A* indicates something to *x*, and if *y* shares *x*'s inductive standards and background information, then *A* must indicate the same thing to *y*. Therefore, if *A* indicates to *x* that *y* has reason to believe that *A* holds, and if *A* indicates to *x* that ____, and if *x* has reason to believe that *y* shares *x*'s inductive standards and background information, then *A* indicates to *x* that *y* has reason to believe that ____ (this reason being *y*'s reason to believe that *A* holds). Suppose you and I do have reason to believe we share the same inductive standards and background information, at least nearly enough so that *A* will indicate the same things to both of us. Then (2) applied to (3) implies:

(4) *A* indicates to both of us that each of us has reason to believe that you will return.

And (2) applied in turn to (4) implies:

(5) *A* indicates to both of us that each of us has reason to believe that the other has reason to believe that you will return.

And so on *ad infinitum*, since each new conclusion begins "*A* indicates to both of us that . . ." Note that this is a chain of implications, not of steps in anyone's actual reasoning. Therefore there is nothing improper about its infinite length. Figure 26 is a more detailed representation of these implications in my case; those in your case could be represented similarly.

Consider next that our definition of indication yields a principle of detatchment: if *A* indicates to *x* that ____ and *x* has reason to

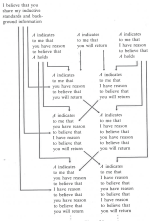


Figure 26

believe that *A* holds, then *x* has reason to believe that ____. Premise (1) applied in this way to (3) implies:

(3′) Each of us has reason to believe that you will return.

Premise (1) applied to (4) implies:

(4′) Each of us has reason to believe that the other has reason to believe that you will return.

Premise (1) applied to (5) implies:

(5′) Each of us has reason to believe that the other has reason to believe that the first has reason to believe that you will return.

And so on, for the whole infinite sequence we considered above. I am still not talking about anyone's actual reasoning or what anyone actually does believe. But the only actual reasoning needed now is reasoning to convert these iterations of "has reason to believe" to the corresponding iterations of "does believe." For that we need ancillary premises about rationality.

Anyone who has reason to believe something will come to believe it, provided he has a sufficient degree of rationality. So according to (3′), if we both have a sufficient degree of rationality, then it will come to be that

(3″) Each of us expects that you will return.

According to (4′), if each of us has reason to ascribe a sufficient degree of rationality to the other, then each has reason to expect that the other expects that you will return. If, in addition, we both have a sufficient degree of rationality, then it will come to be that

(4″) Each of us expects that the other expects that you will return.

According to (5′), if each of us has reason to expect that the other has reason to ascribe a sufficient degree of rationality to him, then

each has reason to expect that the other has reason to expect that he expects that you will return. If, in addition, each of us has reason to ascribe a sufficient degree of rationality to the other, then each has reason to expect that the other expects that he expects that you will return. And if, in addition, we both have a sufficient degree of rationality, then it will come to be that

(5″) Each of us expects that the other expects that he expects that you will return.

And so on. Each term of the sequence (3′), (4′), (5′) . . . , together with sufficient rationality, reason to ascribe sufficient rationality, etc., guarantees formation of the corresponding first- or higher-order expectation. But the degrees of rationality we are required to have, to have reason to ascribe, etc., obviously increase quickly. That is why expectations of only the first few orders are actually formed. The generating process stops when the ancillary premises give out.

This completes our example of a state of affairs which produces higher-order expectations. I take this example to be typical; all the higher-order expectations involved in sustaining conventions, and more or less all we ever have, seem to be produced in this way.

Let us say that it is *common knowledge* in a population *P* that ____ if and only if some state of affairs *A* holds such that:

(1) Everyone in *P* has reason to believe that *A* holds.
(2) *A* indicates to everyone in *P* that everyone in *P* has reason to believe that *A* holds.
(3) *A* indicates to everyone in *P* that ____.

We can call any such state of affairs *A* a *basis* for common knowledge in *P* that ____. *A* provides the members of *P* with part of what they need to form expectations of arbitrarily high order, regarding sequences of members of *P*, that ____. The part it gives them is the part peculiar to the content ____. The rest of what they need is what they need to form *any* higher-order expectations in the way we are considering: mutual ascription of some common inductive standards

and background information, rationality, mutual ascription of rationality, and so on.

Let us return to our example and consider the state of affairs *A* more completely. Suppose that as part of *A* we manifest our conditional preferences for returning to the meeting place. Then *A* may also indicate to us that we both have such preferences. If so, *A* can serve as a basis not only for common knowledge that you will return, but also as a basis for common knowledge that each of us prefers to return if the other does. Suppose also that as part of *A* we somehow manifest a modicum of rationality. Then *A* may indicate to us, and be a basis for common knowledge of, our possession of this modicum of rationality. By now *A*—our incident of agreeing to return—is generating all the higher-order expectations that contribute to our success in solving our coordination problem by means of replication.

A basis for common knowledge generates higher-order expectations with the aid of pre-existing higher-order expectations of rationality. Can these themselves be generated by some basis for common knowledge? Yes, because all the higher-order expectations of rationality needed to generate an *n*th-order expectation are themselves of less than *n*th-order. What cuts off the generation of higher-order expectations is the limited amount of rationality indicated by any basis—not any difficulty in generating higher-order expectations of as much rationality as *is* indicated by a basis.

Agreement to do one's part of a coordination equilibrium is a basis for common knowledge that everyone will do his part. Salience is another basis for common knowledge that everyone will do his part of a coordination equilibrium; but it is a weaker basis, in general, and generates weaker higher-order expectations, since the salience of an equilibrium is not a very strong indication that agents will tend to choose it. Precedents also are a basis for common knowledge that everyone will do his part of a coordination equilibrium; and, in particular, past conformity to a convention is a basis for common knowledge of a tendency to go on conforming. Consider a conventional regularity *R* in a population *P*. Everyone in *P* has reason to

believe that members of *P* have conformed to *R* in the past. The fact that members of *P* have conformed to *R* in the past indicates to everyone in *P* that everyone in *P* has reason to believe that members of *P* have conformed to *R* in the past. And the fact that members of *P* have conformed to *R* in the past indicates to everyone in *P* that they will tend to do so in the future as well.

For example, drivers in the United States have hitherto driven on the right. All of us have reason to believe that this is so. And the fact that this is so indicates to all of us that all of us have reason to believe that drivers in the United States have hitherto driven on the right and also that drivers in the United States will tend to drive on the right henceforth.

Our defining conditions for the existence of a convention consist of a regularity in behavior, a system of mutual expectations, and a system of preferences. I propose to amend the definition: not only must these conditions be satisfied, but also it must be common knowledge in the population that they are. Our amended definition is:

> A regularity *R* in the behavior of members of a population *P* when they are agents in a recurrent situation *S* is a *convention* if and only if it is true that, and it is common knowledge in *P* that, in any instance of *S* among members of *P*,
>
> (1) everyone conforms to *R*;
> (2) everyone expects everyone else to conform to *R*;
> (3) everyone prefers to conform to *R* on condition that the others do, since *S* is a coordination problem and uniform conformity to *R* is a coordination equilibrium in *S*.

Thus there is to be some state of affairs *A* (such that *A* holds, everyone in *P* has reason to believe that *A* holds, and *A* indicates to everyone in *P* that everyone in *P* has reason to believe that *A* holds) which indicates to everyone in *P* that members of *P* conform to *R*, that they expect each other to conform to *R*, and that they have prefer-

ences which make uniform conformity to *R* a coordination equilibrium.

One reason to amend the definition of convention is simply that we want to write into the definition all of the important features common to our examples, and common knowledge of the relevant facts seems to be one such feature. There is another reason: the amendment helps to deal with certain odd cases, regularities which seem intuitively unlike clear cases of convention but which would have qualified as conventions under the unamended definition.

Suppose everyone drives on the right because he expects everyone else to drive on the right and he wants to prevent collisions. But suppose no one gives anyone else credit for intelligence equal to his own. Everyone holds this false belief (call it *f*): "Except for myself, everyone drives on the right by habit, for no reason, and would go on driving on the right no matter what he expected others to do." This is a case of convention under the unamended definition, despite the false beliefs; but I think it ought to be excluded. It cannot be a case of convention under the amended definition (unless we are extremely irrational); for if it is, there is some state of affairs which we have reason to believe holds and which indicates to us that *f* is false. This case is only the first of a sequence. Suppose next that no one really has the false belief *f*, but everyone falsely ascribes it to everyone else. This too cannot be a case of convention under the amended definition (unless we are extremely irrational); for if it is, there is some state of affairs which we have reason to believe holds and which indicates to us that everyone has reason to disbelieve *f*. And so on. The cases become more and more unlikely, but no less deserving of exclusion; the amended definition continues to exclude them (given a sufficiently strong assumption of rationality—stronger and stronger assumptions of rationality are needed as we go on).

By now one might guess that common knowledge is the only possible source of higher-order expectations. But it is not; there is a general method for producing expectations of arbitrarily high order in isolation. For instance, I can acquire an isolated fourth-order

Cubitt and Sugden's formal version of Lewis's analysis

Notation:

- $R_i(p)$: i has reason to believe p.
- $A \operatorname{ind}_i P$: A indicates to i that P
- Cubitt and Sugden's give *four* conditions for A to create common knowledge of P:

For all persons i: A holds $\Rightarrow R_i(A$ holds$)$. (CS1)

For all persons i, j: $A \operatorname{ind}_i R_j(A$ holds$)$. (CS2)

For all persons i: $A \operatorname{ind}_i P$. (CS3)

For all persons i, j:, for all propositions Q: $(A \operatorname{ind}_i Q) \Rightarrow R_i(A \operatorname{ind}_j Q)$. (CS4)

Condition CS4 was implicit in Lewis's text as "suitable ancillary premises regarding our [shared] rationality, inductive standards, and background information".

- Reasons to believe can be arbitrarily nested: $R_i(R_j(\cdots))$.
How can we verify CS4 for all such Q in finite time?

Cubitt and Sugden's formal version of Lewis's analysis

- Neither Lewis nor C&S provide specific semantics for the indication relationship.
- C&S state:

  "Lewis clearly intends **if** … **thereby** … to be stronger than the material implication, $\Rightarrow$. On the most natural reading of the definition of '$A \, \text{ind}_i \, x$', i's reason to believe that A holds provides i's reason for believing that x is true."

- They present six properties that capture their intuition about the requirements for *any* indication relationship.
- We only need two of them (P1 and P6), but I won't discuss these further today.

Cubitt and Sugden's formal version of Lewis's analysis

Their version of Lewis's proof:

Consider any state of affairs A, any proposition P, and any population $\mathcal{P}$. Suppose that A holds and that A is a reflexive common indicator in $\mathcal{P}$ that P. Then:

1.	$\forall i \in \mathcal{P},\ R_i(A \text{ holds})$	(from C1)
2.	$\forall i, j \in \mathcal{P},\ A \,\text{ind}_i\, R_j(A \text{ holds})$	(from C2)
3.	$\forall i \in \mathcal{P},\ A \,\text{ind}_i\, P$	(from C3)
4.	$\forall i \in \mathcal{P},\ R_i(P)$	(from 1 and 3, using P1)
5.	$\forall i, j \in \mathcal{P},\ R_i(A \,\text{ind}_j\, P)$	(from 3, using C4)
6.	$\forall i, j \in \mathcal{P},\ A \,\text{ind}_i\, R_j(P)$	(from 2 and 5, using P6)
7.	$\forall i, j \in \mathcal{P},\ R_i[R_j(P)]$	(from 1 and 6, using P1)
8.	$\forall i, j, k \in \mathcal{P},\ R_i(A \,\text{ind}_j\, R_k(P))$	(from 6, using C4)
9.	$\forall i, j, k \in \mathcal{P},\ A \,\text{ind}_i\, R_j[R_k(P)]$	(from 2 and 8, using P6)
10.	$\forall i, j, k \in \mathcal{P},\ R_i[R_j(R_k[P])]$	(from 1 and 9, using P1)
11.	$\forall i, j, k, l \in \mathcal{P},\ R_i(A \,\text{ind}_j\, R_k(R_l[P]))$	(from 8, using C4)

"and so on"

Our approach and notation (1)

Claim:

To reason about common knowledge, agents need a mechanism for *theory-of-mind* reasoning.

Our approach and notation (2)

- Each agent can choose to maintain a set of named models of other's percepts, beliefs and ToM rules.
- $\odot$ denotes the agent's top-level model.
- *af* denotes "any fool" (McCarthy 1978)[3].
- $\odot \gg af$ denotes the agent's model of any fool's percepts, beliefs and rules.
- $\odot \gg af \gg af$ denotes the agent's model of any fool's model about any (other) fool's percepts, beliefs and rules.
- Example percept and belief propositions:

$$percept(\odot, colour(sky, blue))$$

$$percept(\odot \gg af, colour(sky, blue))$$

$$bel(\odot \gg af, colour(sky, blue))$$

$$bel(\odot \gg af \gg af\ colour(sky, blue))$$

---

[3] In our approach, *af* is a Skolem constant. A set of "*af* scope percepts" can be declared to provide a restricted scope for *af*.

Our approach and notation (3)

- Agents have *theory-of-mind* (ToM) rules that can create new percepts, beliefs and rules in models, e.g.:

  Believe what you perceive

  $percept(\odot, P) \Rightarrow bel(\odot, P)$

  Citizens believe they are citizens

  $bel(\odot, citizen(C)) \Rightarrow bel(\odot \gg C, citizen(me))$

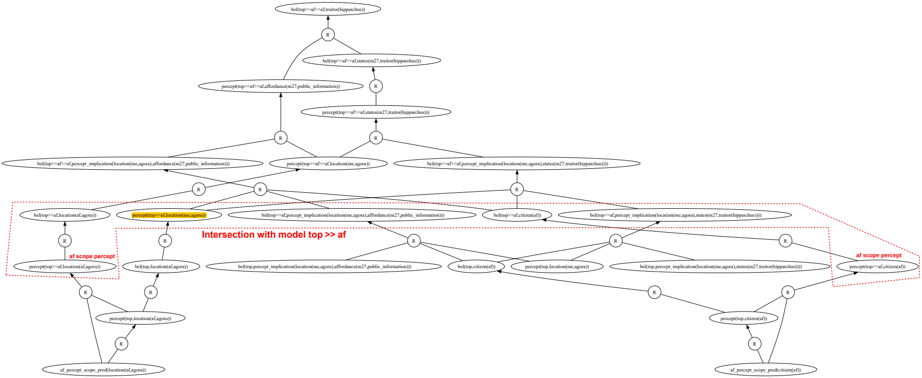
- We interpret states of affairs as sets of percepts.
- We write indication as:

  $percepts(M, A)$ ind ψ

  where ψ is $percepts(M', \cdots)$, $percept(M', \cdots)$ or $bel(M', \ldots)$.
  M' is M or a nested model $M \gg \ldots \gg Ag$.

- We interpret indication as stating that perceiving A in M provides *sufficient* conditions within model M, (in conjunction with the *af* scope percepts), to infer ψ using the ToM rules.

Example proof tree to determine indication



Our versions of conditions C1 to C3

When A is perceived, it is believed that any fool perceives A.

$$percepts(\odot, A) \rightarrow percepts(\odot \gg af, A) \quad \text{(C1)}$$

Believing that any fool perceives A^* is sufficient to infer that any fool believes any fool perceives A.

$$percepts(\odot \gg af, A^*) \text{ ind } percepts(\odot \gg af \gg af, A) \quad \text{(C2)}$$

(A^* is A augmented with the *af* scope percepts.)

Believing that any fool perceives A^* is sufficient to infer that any fool believes P.

$$bel(\odot \gg af, A) \text{ ind } bel(\odot \gg af, P) \quad \text{(C3)}$$

Our version of condition C4

A specialised version of the C&S version: precisely what their proof needs.

$$\begin{aligned}\forall n{\geq}1:\ & \mathit{percepts}(\odot{\gg}\mathit{af}, A) \text{ ind } \mathit{bel}(\odot({\gg}\mathit{af})^n, P)\\ & \rightarrow \mathit{percepts}(\odot{\gg}\mathit{af}{\gg}\mathit{af}, A) \text{ ind } \mathit{bel}(\odot({\gg}\mathit{af})^{n+1}, P)\end{aligned} \tag{C4}$$

This checks that whenever the first indication relationship holds (for any level of nesting n), the equivalent one with an extra "$\gg$ *af*" on each side must also hold.

- **Problem:** This version still cannot be verified using a finite set of ToM models.
- **Solution:** We proved that C4 holds if the models $\odot{\gg}\mathit{af}$ and $\odot{\gg}\mathit{af}{\gg}\mathit{af}$ are *isomorphic*, i.e. they have the same percepts, beliefs and rules (except for the difference in model names).
- **Result:** Only two levels of ToM modelling are necessary to decide whether P is (Lewisian) common knowledge, given a set of percepts A.

Our *inductive* proof that C1–C4 lead to common knowledge of P

Base case

Assumption: $percepts(\odot, A) \xrightarrow{\text{C1}} percepts(\odot \gg af, A)$

C3: $percepts(\odot \gg af, A)$ ind $bel(\odot \gg af, P)$

$\xrightarrow{\text{P1}} bel(\odot \gg af, P)$

Inductive step

$percepts(\odot \gg af, A)$ ind $bel(\odot(\gg af)^n, P)$

$\downarrow$ C4

$percepts(\odot \gg af \gg af, A)$ ind $bel(\odot(\gg af)^{n+1}, P)$

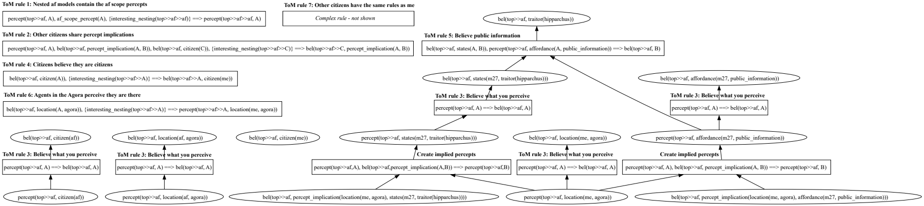
C2: $percepts(\odot \gg af, A^*)$ ind $percepts(\odot \gg af \gg af, A)$

$\xrightarrow{\text{P6}}$ $percepts(\odot \gg af, A^*)$ ind $bel(\odot(\gg af)^{n+1}, P)$

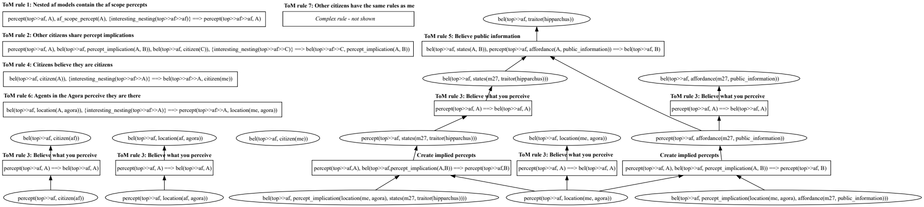
Assumption + af scope predicates : $percepts(\odot, A^*) \xrightarrow{\text{C1}'} percepts(\odot \gg af, A^*)$

$\xrightarrow{\text{P1}} bel(\odot(\gg af)^{n+1}, P)$

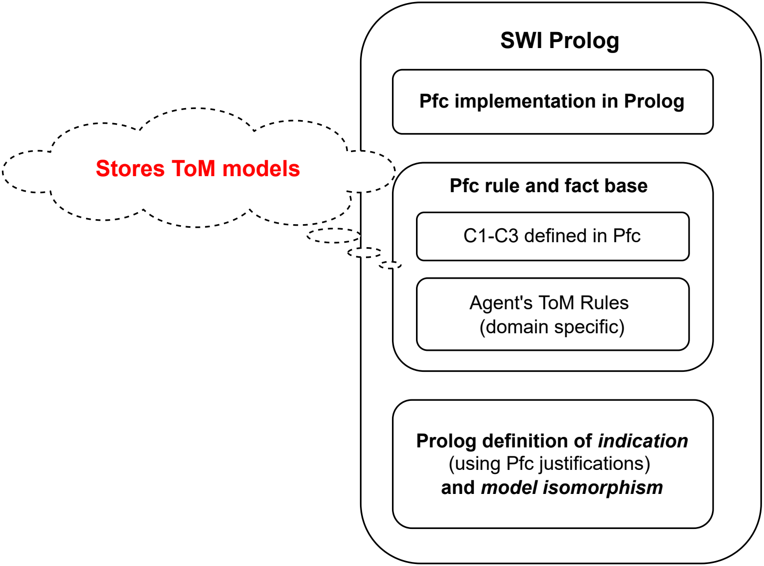
Example model structure comparison for the isomorphism test



⇕ ?



Implementation architecture



Common knowledge conditions as Pfc *backward-chaining* rules

```
c1(A) <==
  { forall(member(Ai, A),
       (percept(top, Ai), percept(top>>af, Ai))) }.

c2(A) <==
  { findall(P, af_scope_percept(P), Ps),
    union(A, Ps, AfAugmentedA),
    percepts(top>>af, AfAugmentedA) ind
      percepts(top>>af>>af, A) }.

c3(A,P) <==
  { percepts(top>>af, A) ind bel(top>>af, P) }.

isomorphic_models(M1, M2) :-
  % Definition in Prolog too long to include

ck(P) <==
  { percepts(top>>af, A) ind bel(top>>af, P) },
  c1(A), c2(A), c3(A,P),
  { isomorphic_models(top>>af, top>>af>>af) }.
```

Example scenario

- Part of a complex prosecutor's argument in a trial for treason in classical Athens (Ober 2010)
  - The prosecutor argued that what happens to traitors is common knowledge.
  - . . . because it is inscribed on a monument in the Agora.
  - How can an agent infer that this is common knowledge?

Scenario ToM rules in Pfc

```
% af scope predicate declarations
==> af_scope_percept(citizen(af)).
==> af_scope_percept(location(af, agora)).

% Create initial af scope percepts:
af_scope_percept(P) ==> percept(top, P).

% ToM 1: Nested af models contain af scope percepts
percept(top, P), af_scope_percept(P),
bel(top, citizen(C)), { interesting_nesting(top>>C) }
==> percept(top>>C, P).

% Percept implication beliefs
==> bel(top, percept_implication(
             location(me, agora),
             states(m27, traitor(hipparchus)))).
==> bel(top, percept_implication(
             location(me, agora),
             affordance(m27, public_information))).

% Create implied percepts
percept(top, P), bel(top, percept_implication(P, Q))
==> percept(top, Q).

% ToM 2: Other citizens share percept implications
percept(top, P), bel(top, percept_implication(P, Q)),
bel(top, citizen(C)), { interesting_nesting(top>>C) }
==> bel(top>>C, percept_implication(P, Q)).

% ToM 3: Believe what you perceive
percept(top, P) ==> bel(top, P).

% ToM 4: Citizens believe they are citizens
bel(top, citizen(C)), { interesting_nesting(top>>C) }
==> bel(top>>C, citizen(me)).

% ToM 5: Believe public information on monuments
bel(top, states(Monument,S)),
percept(top, affordance(Monument, public_information))
==> bel(top, S).

% ToM 6: Agents in the agora perceive they are there
bel(top, location(C, agora)),
{ interesting_nesting(top>>C) }
==> percept(top>>C, location(me, agora)).

% ToM 7: Other citizens have the same rules as me
( Conditions ==> Conclusion ),
{ functor(Conclusion,F,2), memberchk(F, percept, bel),
  conjunction_head(Conditions, Condition),
   ( Condition=percept(M1,_) ; Condition=bel(M1,_) ),
  depth(M1, D), D < 2 },
bel(M1, citizen(C)), { interesting_nesting(M1>>C) }
==>
{ mapsubterms(append(M1,C), Conditions, ModifiedConds),
  mapsubterms(append(M1,C), Conclusion, ModifiedConcl)
},
( ModifiedConds ==> ModifiedConcl ).
```

Example ToM rules in English

- I believe what I perceive.
- Citizens believe they are citizens
- Public information on monuments is believed.
- Other citizens have the same ToM rules as me. This is a *rule-copying* rule, i.e.:

  Given rule $Conds \Rightarrow bel(M, B)$ and $bel(M, citizen(C))$, create the new rule $Conds' \Rightarrow bel(M \gg C, B)$

  where $Conds'$ is $Conds$ with occurrences of M replaced with $M \gg C$.

Inferring common knowledge

- Initial knowledge base
  - Agent's percepts:

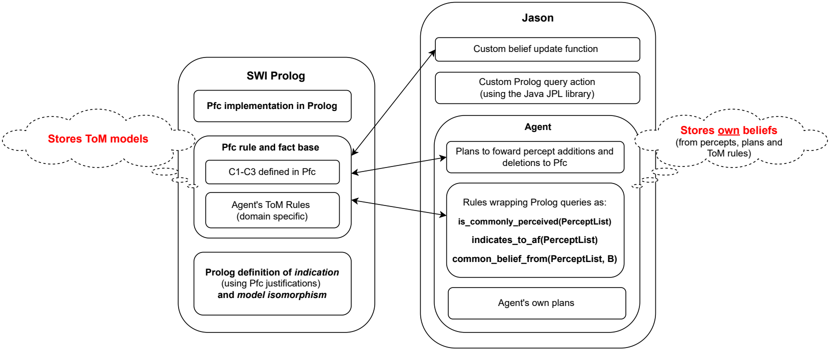
    $percept(\odot, citizen(me))$
    $percept(\odot, location(me, agora))$
    $percept(\odot, states(m27, traitor(hipparchus)))$
    $percept(\odot, affordance(m27, public\_information))$
  - Scope for "any fool":

    $percept(\odot, citizen(af))$
    $percept(\odot, location(af, agora))$
  - Theory of mind rules (Pfc)
  - Rules defining C1 to C3 and overall common knowledge (Pfc)
  - Implementation of indication and isomorphism test (Prolog)
- Query: $ck(P)$
- Result: $P = traitor(hipparchus)$

Prolog query transcript

```
?- bel(top, traitor(hipparchus)).
true.

?- percepts(top>>af, I) ind bel(top>>af, traitor(hipparchus)).
I = [affordance(m27, public_information), states(m27, traitor(hipparchus))] ;
I = [location(me, agora)] ;
false.

?- pfc(c1([affordance(m27,public_information), states(m27,traitor(hipparchus))]
true.

?- pfc(c2([affordance(m27,public_information), states(m27,traitor(hipparchus))]
true.

?- pfc(c3([affordance(m27,public_information), states(m27,traitor(hipparchus))]
true.

?- isomorphic_models(top>>af, top>>af>>af).
true.

?- pfc(ck(traitor(hipparchus))).
true.
```

Integration with Jason (work in progress)



Conclusion

- Agents can coordinate better if they understand what knowledge is common to them all.
- The logic of common knowledge has been investigated for decades, but does not appear to be practically used.
- We adapted Lewis's theory, added the missing ingredient of theory-of-mind reasoning and provided concrete semantics for indication.
- We proved that common knowledge can be inferred with only two levels of ToM reasoning.
- Our approach can be used for agents without rich logical reasoning capabilities, and can be integrated with Jason.
- Talk to me today or at my poster on Wednesday after lunch.

Questions for you

- Have you built agents that reason about common knowledge (or belief)?
- What problem domains do you have where ToM about common knowledge/belief would be useful?