ACTIONS THAT MAKE US KNOW

Johan van Benthem, University of Amsterdam & Stanford University

Revised version, December 2006

Abstract

The knowability paradox is usually formulated as a problem about the static propositions which express the knowledge that we can achieve in principle. In this paper, I propose to put these issues in a more 'dynamic' light, by shifting the emphasis to the *epistemic actions* that produce knowledge, or sometimes even ignorance. The very notion of 'knowability' seems mainly an existentially quantified residue of knowledge-producing actions, just as 'provability' is the static property of propositions that remains when we suppress their live proof and its production. In particular, can every static proposition which is true trigger a dynamic action of announcing that it is true, or of learning that truth? Keeping track of what actions do over time is notoriously difficult, as the truth values of relevant propositions keep changing in processes of computation, physical movement, games, or communication. We discuss some basic issues that arise when we place 'knowability' in a setting of one or more epistemic agents performing a possible variety of epistemic actions.

1 The problem: verificationism incurs the Fitch paradox

Verificationism is an account of meaning and truth whose origins lie in logical proof theory, especially, in its constructivist versions. The idea is that 'truth' can only be assigned to propositions for which we have evidence. This view is found with logical authors like Dummett and Martin-Löf since the 1970s, but it has also penetrated since into general philosophy. Stated as a sweeping claim, this take on truth implies the general verificationist thesis that *what is true can be known*:

$$\phi \rightarrow \Diamond K\phi \qquad \textbf{\textit{VT}}$$

Here the K can be taken as a relatively unproblematic knowledge modality, while the $\Diamond$ is an as yet unspecified modality "can" of 'feasibility' in some relevant sense. Now, a surprising argument by Fitch trivializes this principle. It uses just a weak modal epistemic logic to show that ***VT*** collapses the notions of truth and knowledge, by taking the following clever substitution instance for the formula ϕ:

$$q \wedge \neg Kq \rightarrow \Diamond K(q \wedge \neg Kq)$$

Then we have the following chain of three conditionals – which works in quite weak and apparently unproblematic modal logics:

$$\Diamond K(q \wedge \neg Kq) \rightarrow \Diamond (Kq \wedge K\neg Kq)$$
$$\rightarrow \Diamond (Kq \wedge \neg Kq) \rightarrow \Diamond \bot \rightarrow \bot$$

Thus, a contradiction follows from the assumption $q \wedge \neg Kq$, and we have shown overall that q implies Kq, making truth and knowledge equivalent.

Is there a real problem here? How plausible was Verificationism anyway? There can be legitimate doubt on this score – but all the same, looking at 'paradoxes' like Fitch's can be worth-while. Of course, not every paradoxical argument points at a genuine problem. Some are just spats on the Apple of Knowledge, which can be removed with a damp cloth. But others are the tell-tale brown spots of worm rot inside, and deep surgery is needed – and the Apple may not even remain in one piece when restoring consistency. Professional paradox hunters and puzzle-driven researchers always claim the 'deep trouble' diagnosis of course – and sometimes they are even right.

Proposed remedies for the Paradox so far fall mainly into two kinds (cf. Brogaard and Salerno 2002, van Benthem 2004). Some solutions weaken the logic in the argument still further. This is like tuning down the volume on your radio so as not to hear the bad news. You will not hear much good news either. Other remedies leave the logic untouched, but weaken the verificationist principle itself. This is like censoring the news: you hear things loud and clear, but they may not be so interesting. Some choice between these strategies is inevitable. But what one really wants is a *new systematic viewpoint* going beyond plugging holes, and opening up a new line of thinking with benefits elsewhere. In our view, the locus for this is not Fitch' proof as such, but rather our understanding of the two key modalities involved, either the K or the $\Diamond$, or both.

2 A first quick analysis: epistemic logic and evidence

Let us first get to the essence of Fitch's argument. The above substitution instance exemplifies a much older problem called *Moore's Paradox*. Originally stated about belief, it consists in the observation that the statement

"*P*, but I don't believe it"

can be true, whereas it cannot be consistently believed. Transposed to knowledge, this same problem was discussed by Hintikka in the 1960s, using the inconsistency of the formula $K(q \,\&\, \neg Kq)$ in epistemic logic. So, it is easy to understand the issue. Some truths are 'fragile' whereas knowledge is 'robust': and hence the former need not support

the burden of the latter. Thus, one sensible and straightforward approach to the paradox weakens the scope of applicability of *VT* as follows (Tennant 2002):

Claim *VT* only for propositions ϕ such that $K\phi$ *is consistent* ***CK***

CK has clear merits, but it fails our more general desideratum: it provides no exciting new account of either knowledge *K* or feasibility $\Diamond$. We have put our finger in the dike, but no larger polder management system has emerged. Indeed, there seems even an obvious missing link in ***CK***, reflecting one's intuitive semantic understanding of the setting for *VT*. We have the truth of ϕ *in some epistemic model* ***M*** with actual world *s*, representing our current information state. But consistency of $K\phi$ per se gives us only the truth of $K\phi$ in some possibly *quite different* epistemic model *(**N**, t)*. The issue is:

What natural step of 'coming to know' would take us from *(**M**, s)* to *(**N**, t)*?

One could see this as asking for a principled account of the above operator $\Diamond$, while the *K* can retain its standard meaning from epistemic logic.

One way in which the $\Diamond$, has been unpacked in the literature goes back to the proof-theoretic origins of *VT*. In well-established type-theoretic approaches to provability, the *evidence* for a conclusion is displayed and manipulated in binary assertions of the form

$p: \phi$, where *p* is a proof for ϕ, or a piece of evidence in a more general sense.

Type theory seems the most sophisticated underpinning of Verificationism to date. Van Benthem 1993 took this idea to standard epistemic logic, and proposed an explicit calculus of evidence for its *K*-assertions. One striking modern realization of this is the 'logic of proofs' of Artemov 1994, 2005, which replaces the box $[]\phi$ of the usual modal provability logic by operators $[p]\phi$: '*p* is a proof for ϕ'. Indeed, labels *p* of many sorts appear in the 'labeled deductive systems' of Gabbay 1996. This 'evidence parameter' for logical investigation seems a deep response to any paradox – but I am not aware of an inspiring solution to Fitch-style problems in this proof-theoretic setting.

Thus, I take a different tack here, in terms of semantic actions that produce knowledge.

3 Dynamics of information and coming to know

Broadening our view of what a feasibility modality $\Diamond$ might stand for, van Benthem 2004, 2006A look at general mechanisms producing knowledge. Mathematical proof, no matter how liberally construed, is not the best paradigm for understanding how we come to know things, since it does not add new truths beyond our premises. Genuine actions by which we come to know new things seem much more domestic: we *observe*,

or we *ask* some expert who knows! The latter actions involve a notion of change beyond proof steps: new information changes the current epistemic model – and in the process our knowledge changes, too. The simplest mechanism achieving this reflects the folklore sense in which 'new information shrinks the current range of possibilities':

> An announcement of some proposition *P* changes the current range of possible worlds, leaving those where *P* holds, while removing all others.

More precisely, consider an epistemic model $(\mathbf{M}, s)$, with designated actual world s. What can be known in this setting seems restricted to what might be known correctly *about that actual situation s*. We know already that it is one of the worlds in $\mathbf{M}$. What we might learn is that this model can be shrunk further, zooming in on the location of s. In this dynamic epistemic setting, we can recast the Verificationist Thesis as follows. Saying that every true statement may be known amounts to stating that there is

What is true in the current setting may come to be *known there* **VT-dyn**

What this means in a simplest scenario is that some authoritative true statement could be made which *changes the current model* $(\mathbf{M}, s)$ to some submodel $(\mathbf{M}|\phi, s)$ where the relevant proposition ϕ is known. Indeed, *announcing* ϕ *itself* seems an obvious and infallible candidate for this purpose, but more on this in a moment.

The dynamic turn toward knowledge-producing actions involves some delicate issues. A first thing to note is that making announcements is not just a matter of accumulating knowledge. This is true for atomic facts – but truth values of more complex epistemic assertions can change in the process. When I tell you that p, which you did not know, the statement $K_{you}p$ changes its truth value from false to true. But at the same time, the iterated knowledge statement $K_{you}\neg K_{you}p$ goes from true to false – and so on upward, with changes in iterated statements of epistemic reflection. Thus, one single action

$!\phi$

of publicly announcing ϕ can have repercussions for truth values across the epistemic language. In particular, the Moore sentence shows that some propositions ϕ have the 'self-afflicting' property of changing their *own* truth value when they are announced:

> A true public announcement $!(q \,\&\, \neg Kq)$ of $q \,\&\, \neg Kq$ makes the fact q into common knowledge, thereby invalidating the conjunct $\neg Kq$.

Thus, announcing a truth is not an infallible way of turning it into knowledge. We will investigate the subtleties of epistemic update in the next section. For now, we contrast our new dynamic view with the earlier consistency requirement on *CK*.

Here is the connection between our new proposal ***VT-dyn*** and the earlier ***CT***:

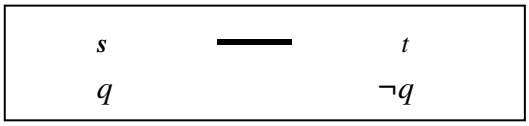
Fact (a) ***VT-dyn*** implies ***CK***

(b) ***CK*** does not imply ***VT-dyn*** for all propositions ϕ.

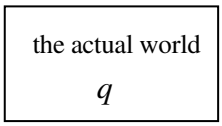
Proof Implication (a) is obvious. Its converse (b) is not, as we need truth of $K\phi$ not in just in any model (which would suffice for consistency), but *in some submodel of the current one*. Here is a counter-example. Not surprisingly by now, it works with a relative of the Moore-type assertion $q \,\&\, \neg Kq$:

$\phi = (q \,\&\, \neg Kq) \vee K\neg q$ where q is a proposition letter.

This is knowable in the sense of ***CK***, since $K((q \,\&\, \neg Kq) \vee K\neg q)$ is consistent. For instance, this formula holds in a model consisting of just one world with $\neg q$. Indeed, in S5, the statement $K\phi$ is equivalent to $K\neg q$. But now consider the following two-world epistemic *S5*-model ***M*** with an actual world ***s*** and an epistemically indistinguishableworld *t*, where the atomic formula q holds at *s* but not at *t*. In this situation, no truthful announcement would ever make us learn the above ϕ:



In the actual world, $(q \,\&\, \neg Kq) \vee K\neg q$ holds, but it fails in the other one. Hence, $K((q \,\&\, \neg Kq) \vee K\neg q)$ fails in the actual world. Now, there is only one truthful proper update of this epistemic model ***M***, which just retains its actual world with q:



But in this one-world model, the formula $K((q \,\&\, \neg Kq) \vee K\neg q)$ evidently fails. ♥

The preceding example suggests that ***CT***, though correct in spirit, is still too weak in a dynamic setting. This point is somewhat technical, but telling all the same. It shows how, in a natural semantic scenario of coming to know things, the Verificationist Thesis places stronger requirements on propositions than those in the literature so far.

How can this happen? Why does not a true assertion $(q \,\&\, \neg Kq) \vee K\neg q$ stay true when we 'learn more'? Once again, the learning intuition behind world elimination is only valid for *factual propositions*. But *epistemic propositions* involving *K*-modalities may change their truth value when a model contracts, as ignorance has now turned into

knowledge. To understand this better, let us now look in more detail at logical mechanisms for epistemic change and learning (van Benthem 2002, 2006A, 2006B).

5 Epistemic logic dynamified

Static epistemic logic The basic language of epistemic logic and its semantics are well-known, with the individual knowledge modality $K_i\phi$ interpreted as follows:

> $K_i\phi$ is true at a world s iff ϕ is true in all worlds t with $s \sim_i t$,
> where $\sim_i$ is the epistemic accessibility relation for agent i.

In what follows, for convenience of exposition, we use an *S5* version, where world accessibility is an equivalence relation. This simple semantics of knowledge has inspired much philosophical discussion, partly by the logical precision that it offers, but also by its perceived deficiencies. Hotly debated until today are 'logical omniscience' (closure of knowledge under valid implications), and 'introspection' (automatically knowing that one knows or does not know a proposition): cf. van Benthem 2006A.

Moving beyond single agents, epistemic logic can also analyze new forms of 'social' knowledge in groups. In particular, *common knowledge* $C_G\phi$ for a group G says intuitively that everyone knows that ϕ, they also know that the others know, and so on to any finite depth of iteration of mutual knowledge operators. Semantically, the corresponding epistemic modality

> $C_G\phi$ is true at a world s whenever ϕ is true in the whole
> 'component' of the model consisting of all worlds accessible
> from s by some finite sequence agent accessibility steps.

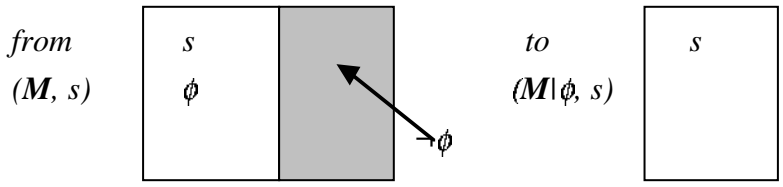
In scenarios with just a single agent *1*, common knowledge $C_{\{1\}}\phi$ is just the same as knowledge $K_1\phi$. (This would not work in weaker epistemic semantics than that for S5.) One can read the following discussion up to Section 7 either way, as being about knowledge of a single agent, or about common knowledge in a group.

As for epistemic inference, well-known complete axiom systems exist for the valid laws in this language over standard model classes, such as multi-agent *S5* (plus common knowledge) for models where the accessibilities are equivalence relations. Finally, as to computational complexity, most current versions of epistemic logic are decidable.

Dynamic epistemic logic To deal with the dynamics of Section 4, we need to add epistemic actions to this framework. Here, the driving engine for update of agents' information is *model change*. The simplest case described earlier is that of a truthful

public announcement $!\phi$ of an assertion ϕ.

This does not just evaluate ϕ truth-conditionally in the current model $(\boldsymbol{M}, s)$. It rather updates that model to a new model $(\boldsymbol{M}|\phi, s)$, a submodel of $(\boldsymbol{M}, s)$ – and it does so by eliminating all those worlds from it which fail to satisfy ϕ:



This update scenario can analyze questions and answers producing new information, and it even works for much more intricate puzzles involving knowledge and ignorance (van Benthem 2002). Thus, we get a *dynamic-epistemic logic*, as a general semantic setting for information flow and learning. There is a family of epistemic models: the relevant information states, and a repertoire of announcement actions, which increase information by moving from one model to another. Full-fledged dynamic-epistemic logics arise from standard epistemic ones by adding an action modality from dynamic logics of computation. It expresses what holds after an action was performed:

$\boldsymbol{M}, s \models [!\phi]\psi$ iff $(\boldsymbol{M}|\phi, s) \models \psi$

Thus, the dynamic modality $[!\phi]\psi$ says that "after ϕ has been truthfully announced, ψ holds at the current world". With this language, one can express systematic effects of communication, using combined knowledge-action statements such as

$[\phi!]K_j\psi$ after a public announcement of ϕ, agent j knows that ψ

There are complete and decidable logical calculi for this richer language, too. Their key axioms analyze the result of an epistemic action in terms of things that were true before.

Dynamic epistemic logic does not magically solve the problems of static epistemic logic, as perfect reasoning with logical omniscience, and perfect reflection with epistemic introspection are still assumed. But our new logics do help analyze and even high-light further issues of potential philosophical interest. Sometimes, it is just liberating to move to new problems instead of remaining stuck with old ones. In particular, one additional idealization of the dynamic setting seems worth pondering. The central valid law of the logical calculus of public announcement reduces knowledge resulting from communication to *relativized* knowledge that was true before:

$[A!]\, K_i\, \phi \;\leftrightarrow\; (A \rightarrow K_i\, (A \rightarrow [A!]\, \phi)$.

The semantic soundness of this principle has its own further presuppositions, including *perfect memory* of agents (Liu 2006). This idealization has been called into question in game theory and cognitive psychology under the heading of 'bounded rationality'.

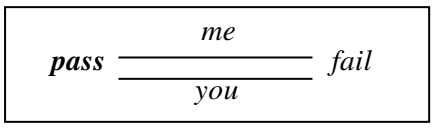
Moving beyond single knowers, however, the most exciting applications of epistemic logics to-day emphasize the *multi-agent* character of speakers, hearers, and audiences. In particular, even Hintikka's original language can iterate knowledge assertions, as in

$K_1 \neg K_2 P$ "*1* knows that *2* does not know that *P*"

Also, common knowledge was a group phenomenon par excellence. 'Social' epistemic notions are crucial to information flow and communication. Some philosophers think such issues are not profound, having to do with gossip, *ICT*, and other shallow necessities of living with a lot of people on one small planet. But the pursuit of knowledge and rational behaviour consists to a large extent of *intelligent interaction* with others – and we need to understand that success. This point will return below, as so-called paradoxes afflicting lonesome knowers may look brighter in groups.

Interaction, partial observation, and event update Public announcement is a basic mode of transmitting information. But information can flow in many more subtle ways. E.g., we observe informative events without overt linguistic aspects. And crucially, observation can then be different for different observers. I see *which* card I am drawing from the current stack, you only see *that* I am drawing one. By now, sophisticated *event update* mechanisms exist for such phenomena, far beyond simple world elimination (Baltag, Moss & Solecki 1998). These can model complex multi-agent forms of communication mixing public actions and *information hiding*. Think of whispering to your colleagues during a seminar, or sending an email using the button *bcc*. In cases like these, the current epistemic model need not shrink: it may even grow in size.

Example: Reading a Letter You have taken an exam, but neither you nor your friend knows the outcome yet. Here is a simple epistemic model, where in fact (viz. the bold-face actual world to the left), you passed:



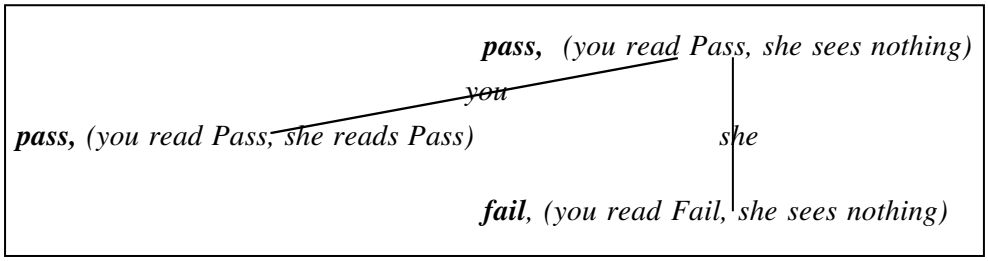
Now you receive a letter in the presence of your friend, and read that you have passed. If this were a case of public announcement, the model would just shrink to the left-hand world as before. But this time, you cannot tell whether your friend has seen the content of the letter, though she does know it is an official notification. She might, and she might not have seen what you read – and so, as far as she is concerned, you might also

have been reading a letter which says that you failed. In this case, taking both your situations into account, there are *3* relevant possible pairs of simultaneous events:

(you read Pass, she reads Pass)
(you read Pass, she sees nothing)
(you read Fail, she sees nothing)

The first two joint events can only occur if you have passed, the third if you failed. Note also that these events themselves have epistemic relations. E.g., you cannot distinguish the first from the second, and she cannot distinguish the second from the third.

Next, as for update, the new epistemic model resulting from incorporating the new information in the reading/observing event into the preceding two-world model has *3* instead of *2* worlds now, with pairs *(old world, new event)* as depicted here:



Here the new epistemic relations arise as follows. Agents cannot distinguish two pairs *(s, e)* and *(t, f)* if they can distinguish neither the old worlds *s, t* nor the new events *e, f*. Suppose that in fact your friend read what was in the letter. Then the actual world is

pass, *(you read Pass, she reads Pass)*

In that world, by standard evaluation in terms of epistemic logic, you know that you passed, she knows it, too, but you do not know that she knows. These are typical asymmetries of information that arise between players in the course of a card game.

General event update takes a model ***M*** for the current information of a group of agents, plus some event model ***A*** modelling all relevant events, and then compute a new

'product model' ***MxA***.

This construction covers much of the information flow in communication, games and other more realistic activities. Again, there are complete and decidable dynamic-epistemic logics dealing with what agents know stage-by-stage as general actions of this sort take place (van Ditmarsch, van der Hoek & Kooi 2006).

6 Learning by update

The self-refuting nature of true Moore-type assertions noted in Section 4 shows that Fitch-style issues about Verificationism reflect core phenomena in information update. Indeed, this analogy is the main point of this paper. But it is of interest to see how these issues play in dynamic epistemic logic. They do not have the same doom-laden atmosphere. Informal studies of speech acts sometimes state that the generic effect of a public announcement $!\phi$ is simply that ϕ becomes common knowledge. The neat corresponding axiom in the earlier calculus would read as follows:

$$[!\phi]C_G\phi$$

But this principle is not valid in general, witness the earlier-mentioned Moore sentence $\phi = q \;\&\; \neg Kq$. The latter assertion, once announced, cannot be true any more – and it even makes its own negation common knowledge! The reason is that announcing ϕ makes q common knowledge, and hence also Kq, but Kq implies $\neg(q \;\&\; \neg Kq)$.

This is not an isolated curiosity. Gerbrandy 2007 gives a new analysis of the well-known Paradox of the Surprise Examination, which revolves around a teacher's problematic assertion that some upcoming exam in the following week will take place on a day 'when the student does not expect it'. Gerbrandy shows how the usual perplexity dissolves once we see that the teacher's assertion can be of the above true–but–self-refuting type. E.g., with a two-day time span, the formula for the teacher's statement in our dynamic-epistemic logic is this (writing *Ei* for 'the exam is on day *i*'):

$$(E1 \;\&\; \neg K_{you}\, E1) \vee (E2 \;\&\; [!\neg E1]\neg K_{you} E2):$$

This says that the exam is on Day 1, and you do not know that now, or it will be on Day 2, and even learning that it is not on Day 1, you will not know that it is on Day 2. For details and a further defense of this analysis, we refer to the cited publication. Simple epistemic models of the above sort then clarify various surprise exam scenarios.

From paradox to typology These observations do not suggest at all that one must *ban* self-refuting assertions – as has been proposed in some remedies to the Fitch Paradox. To the contrary, they rather bring to light a rich diversity of types of behaviour which calls for a *dynamic typology of epistemic assertions*. E.g., we can investigate *which precise* forms of assertion are 'self-fulfilling', in that they *do* become common knowledge upon announcement. For instance, all *universal* modal formulas are self-fulfilling in this sense. These are the ones constructed using

atoms and their negations, conjunction, disjunction, K_i and C_G.

But there are other self-fulfilling types of statement, and a complete syntactic characterization has been an interesting open model-theoretic problem since the late 1990s (Gerbrandy 1999, van Benthem 2002, van Ditmarsch & Kooi 2006).

Technical logical studies in this vein have brought to light further delicate phenomena. In particular, some epistemic assertions ϕ are only self-fulfilling or 'self-refuting' *in the long run*. When announced truly for as long as possible, they either result in common knowledge $C_G\phi$, or the opposite: $C_G\neg\phi$. Van Benthem 2002 applies this insight to game theory, and shows how well-known game solution procedures may be analyzed in terms of repeated announcement of formulas ϕ expressing the 'rationality' of all players. Such statements are informative in general, and remove possible strategic equilibria, but at the first stage where they no longer shrink the model, common knowledge of rationality sets in. Thus, instead of exorcising paradox, we chart the diversity of epistemic behaviour. This turn may be compared to that in Kripke's theory of truth, where self-reference became an object of study, rather than a taboo.

Another interesting typology goes back to the 'coming to know' of Section 4, our dynamic setting for learning true propositions. Indeed, van Benthem 2004 defines three possible types of learnability for propositions ϕ, using an existential action modality

$<!A>\psi$ 'one can truly announce A and then get ψ true':

$\models \phi \rightarrow \exists A\, <A!>K\phi$	*Local Learnability*
$\exists A: \models \phi \rightarrow <A!>K\phi$	*Uniform Learnability*
$\models \phi \rightarrow <\phi!>K\phi$	*Autodidactics*

He shows that each successive type is more demanding than the preceding. Moreover, at least on epistemic *S5*-models, all three notions of learnability are decidable. Further notions of learning arise with iteration of true assertions, perhaps even the same one. Baltag, van Ditmarsch, Herzig, Hoshi & de Lima 2006 present sophisticated update calculi of this sort, and they prove in particular that, when added to our basic logic of public announcement, the logic of 'truth *after some announcement'* stays axiomatizable.

Thus, once again, the 'paradox of knowability' turns from a nuisance into an interesting phenomenon to be studied, and a source of intriguing new logical questions.

Digression: reachability with event updates The event updates of Section 5 took an epistemic model $\boldsymbol{M}$ and an event model $\boldsymbol{A}$, and computed a new product model $\boldsymbol{MxA}$. Many more statements may be made true by such drastic changes. Call a model $\boldsymbol{N}$ 'reachable' from $\boldsymbol{M}$, if, for some event model $\boldsymbol{A}$, $\boldsymbol{N}$ is equal (or better: 'epistemically bisimilar') to $\boldsymbol{MxA}$. Could this approach rescue *CT*? We gave a model $(\boldsymbol{M}, w)$ and a

statement ϕ true in it – but, even though $K\phi$ was consistent, having a model $\boldsymbol{N}$ with ϕ true throughout, no eliminative update took $\boldsymbol{M}$ to such a model. But could some more general event update do that job? For the *single-agent* case, this is not so. Every model $\boldsymbol{MxA}$ is then bisimilar to some submodel of $\boldsymbol{M}$. But, there might be another way of saving *CT*. To link with the current $(\boldsymbol{M}, w)$, we might require that $K\phi$ be consistent with a description of the current world in $(\boldsymbol{M}, w)$. But, if we make $K\phi$ consistent with the *state description* of w (its true and false atomic propositions), update may still be impossible. If we make $K\phi$ consistent with the *complete modal theory* of w, we do get a model bisimilar to $(\boldsymbol{M}, w)$ (van Benthem 2002), but this seems a trivial victory. ♣

Explicit temporal perspectives on knowledge We conclude with a common criticism of the 'learning problem' in the dynamic epistemic setting. Self-refuting Moore-type assertions evoke strong responses. One either loves this sort of subtlety, or one thinks it fundamentally misguided, disregarding the role of *time*. And indeed, there is a sense in which announcing any true proposition *should* always lead to common knowledge. When I say that ϕ is true right now, at time t_0, immediately afterwards, it becomes common knowledge that ϕ *was* true *then* at time t_0! This insight is not in conflict with the type of logic we have used. We can add explicit temporal operators to the dynamic epistemic framework, say a Y for *yesterday* in the time of our epistemic process. Then we get a complete update logic again, including the following attractive validity:

$$\phi \rightarrow [!\phi]C_G Y\phi$$

This says that, if ϕ *is* true now, announcing it makes it common knowledge that it *was* true *at the preceding stage*. Some conversational moves work in just this way – like when people say in response to some assertion that "I knew already what you told me". One might see this as one plausible sense in which the Verificationist thesis does hold:

> Every local truth *right now* can come to be known as
> *being true now* at some later stage of investigation.

Indeed, analyzing the Paradox of Knowability in an explicit temporal epistemic logic has been proposed before, e.g., in Edgington 1985. In such a formalism, all the above issues still make sense. In particular, we now want to know precisely which assertions will persist over time, from $Y\phi$ to ϕ. For some further explorations at the border-line with dynamic-epistemic logic, cf. Sack 2006, Yap 2006, van Benthem & Pacuit 2006. It has to be said that this greater expressive power also has its price. In particular, statements of valid 'learning principles' , and complexity of epistemic-temporal logics, depend in subtle ways on which precise strength we give to the temporal operators.

This section has presented a number of technicalities that may seem non-germane to our general discussion. But the way we see it, these demonstrate that any 'banning' response to the Fitch paradox would be a bad idea, as it would deprive us of a rich area of investigation, offering a lot of genuine insight into how we come to know things.

7 Many agents, communication, and interaction over time

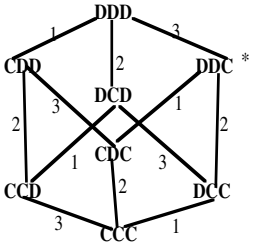
Modern epistemic logic is no longer about lonely knowers in rickety armchairs in leaking attics. It unfolds its true attractions in multi-agent settings, analyzing what agents know about each other, and how they interact: in communication, games, or any other social activities where information flows. The earlier-mentioned notion of common knowledge is crucial then. Here, self-refuting assertions come up naturally without any Moore-like paradoxical flavour, witness this evergreen from the literature:

Puzzles of repeated announcement Like other areas of logic, dynamic epistemic logic has its 'icons'. In the well-known puzzle of the Muddy Children, whose epistemic importance was recognized in Fagin, Halpern, Moses & Vardi 1995, it is successive public announcements of ignorance which drive the solution process toward common knowledge of the true state of affairs. In a simple version, the story runs as follows:

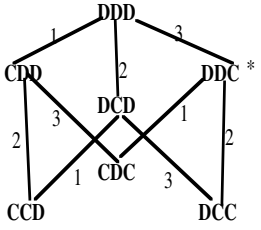
> *After playing outside, two of three children have mud on their foreheads. They all see the others, but not themselves, so they do not know their own status. Now their Father comes and says: "At least one of you is dirty". He then asks: "Does anyone know if he is dirty?" The children answer truthfully. As questions and answers repeat, what will happen?*

Nobody knows in the first round. But in the second round, each muddy child can figure out her status, by explicit reasoning, or by updates. To display these, draw an epistemic model whose worlds assign ***D*** or ***C*** to each child. The actual world is ***DDC***: that is, child *1* and *2* are dirty, while child *3* is clean. Initially, a child knows only the status of the others' faces, but not her own The corresponding epistemic uncertainty relations are indicated by the labeled lines in the following diagrams. Epistemic updates start with the Father's elimination of the world ***CCC***:

from

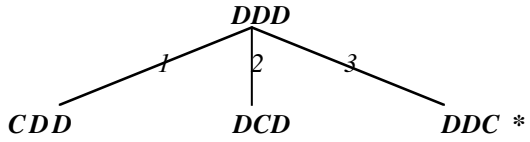


to



One can see this as a simple 'symmetry breaking' of the original pattern which will have

startling consequences – like the way, say, a professional starts a snooker game. Next, when it turns out that no one knows his status, the bottom worlds disappear:

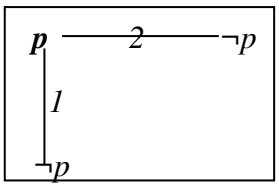


Finally, when the muddy children *1, 2* say simultaneously that they know their status, all worlds where at least one of them still has an uncertainty line left disappear. Thus, this statement, too, was highly informative – and the final update is to

DDC *

With *k* muddy children, *k* rounds of public ignorance assertions achieve common knowledge about who is dirty, while the announcement that the muddy children know their status achieves common knowledge of the whole situation. Thus, public assertions of ignorance can drive a positive process of gathering information, and their ability to eventually invalidate themselves (the earlier-mentioned phenomenon of 'self-defeating' assertions) may even be the crowning event. The last announcement of ignorance for the muddy children led to their knowing the actual world. This puzzle high-lights the interplay of many agents, and also the passage of time. We consider both in turn.

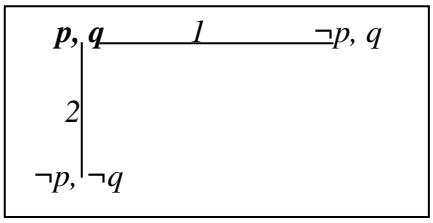
Multi-agent learning Our scenario suggests that learning becomes more interesting, and less 'paradoxical' in a multi-agent setting. Indeed, we do not need Muddy Children to make this point. Even much simpler epistemic models can represent interesting scenarios of communication which might be hard to keep straight just in words. Consider the following simple example involving three worlds and two agents:



In the actual world to the top left, indicated in black, *p* is in fact the case, but neither agent knows if *p*. Now what can the agents learn by internal communication? First, neither can tell the other something factual about *p*. And yet the agents can discover where they are by *communicating their epistemic state*, as follows. First *1* says *"I don't know if p"*. This rules out the right-most world, where $K_1\neg p$ holds. After the update, only the left-hand worlds remain, and so *2* now knows that *p*. Saying thát will then also inform *1*, and the agents have achieved common knowledge that *p*. This is just one example: even three-world models support a whole range of communication scenarios.

Thus, epistemic models with different accessibility structure for agents have a potential for useful information exchange, either of ground facts, or of epistemic attitudes.

The more general issue here is what agents can learn if they communicate what they know, and keep doing so until the model no longer shrinks. Van Benthem 2002 describes such scenarios completely, showing they lead to a unique submodel (with the technical proviso of 'modulo bisimulation'). Essentially, internal communication turns the 'implicit knowledge' of a group into common knowledge. Similar scenarios have been studied in game theory when making hidden correlations in information explicit between players. But there may be other, more complex sorts of specification for a communication process . E.g., we may want only some group members to learn that ϕ, while keeping the others in the dark. This can also happen with Moore-type statements. Here is one more concrete scenario showing such multi-agent phenomena. Consider the following model ***M*** with actual world ***p, q***:



Announcing *q* will make *2* know the Moore statement that "*p* and *1* does not know it". But this can never become common knowledge in the group *{1, 2}*. What can become common knowledge, however, is *p & q*, when *1* announces that *q*, and *2* then says *p*.

Fitch paradoxes for plural knowers? Next, consider the Paradox of Knowability once more. When *q* is true and you don't know it, there is nothing problematic with *others* knowing both those facts. Indeed, general communicative actions – though not full public announcement to the whole group – can ensure the truth of

$K_2(q \,\&\, \neg K_1 q)!$

Thus we might amend the Verificationist Thesis *VT* once more, and recast it as:

If ϕ is true, then *someone* could come to know it. $VT_{multi\text{-}agent}$

This principle is true, at least in some construals! In any model $(\boldsymbol{M}, s)$ where ϕ holds, adding a perfectly informed agent whose epistemic accessibility relation is identity between worlds is consistent, and in the expanded model that new agent knows that ϕ.

More interesting seems the issue of coming to know facts about the whole group. Here are two new possibilities. First, let ϕ be true but not common knowledge:

$\phi \,\&\, \neg C_G \phi$

This cannot be common knowledge in the group G, as the old Fitch argument still applies. But $\phi \,\&\, \neg C_G \phi$ *can* be known by individual agents, and even whole subgroups. Next, consider a stronger case: ϕ is true, but there is a *false common belief* that it is not:

$\phi \,\&\, CB_G \neg \phi$

This time, no agent in the group can come to know this – at least in a very plausible epistemic-doxastic logic.. For if agent i were to know $\phi \,\&\, CB_G \neg \phi$, we would have

(a) $K_i\phi \rightarrow K_iK_i\phi \rightarrow K_iB_i\phi$, (b) $K_iCB_G\neg\phi \rightarrow K_iB_i\neg\phi$, and so
(c) $K_i(B_i\neg\phi \,\&\, B_i\neg\phi)$, and $K_iB_i\bot$, and hence a contradiction,
at least, if our logic does not allow belief of contradictions.

Thus, Verificationism becomes a more varied issue in communities of epistemic agents.

Temporal perspective once more: game theory and learning theory Dynamic epistemic logics describe single steps in larger processes where information flows. There seems to be a growing consensus that such long-term procedures are crucial to 'coming to know'. Our concerns so far then merge into larger issues about *interactive agents* with goals and strategies for achieving them. Thus, dynamic-epistemic logic meets *game theory* (Osborne & Rubinstein 1994) and *learning theory* (Kelly 1996), including strategic equilibria and convergent learning procedures in both finite and infinite settings. These links go beyond the present paper, but their import is clear. In the final analysis, *what* one can come to know is intimately intertwined with the *how*!

8 Conclusion

We have looked at the Paradox of the Knower in a dynamic-epistemic perspective where learning means changing the current epistemic model. The problematic Moore sentence driving the paradox turns out to be the typical 'probe' for investigating the sometimes surprising, but always useful, effects of successive assertions. Moreover, the multi-agent setting of epistemic logic places Verificationism in a richer interactive setting. This change in perspective trades the atmosphere of paradox and disaster for one of free exploration of dynamic typology of epistemic assertions, learning and reachability, and many further surprising twists in the logic of communication.

Even so, we do not claim the last word on Verificationism, the origin of the Fitch puzzle. The proof-theoretic paradigm of *evidence* for what we know also has a ring of truth. And indeed, the dynamic approach so far has no insightful take on the 'information' that comes to us via deduction (cf. Egré 2004, Jago 2006). Updating with

logical consequences of what we know does not change any of the models used here. A unified account of learning from deduction and from observation is a long way off.

And even our own semantic perspective has told only half of the story. In a truly multi-agent setting, learning is not a single-agent matter, and the basic paradigm should have at least two roles: the *Learner* and the *Teacher*. And then, the issue with learning is not just what information we get when updated by some given assertion – say, an answer – or some more general observation of an event. It is just as much the other side of the coin: what we ask of others, and how we enquire. Verification and verificationism seems really about both *seeking* and *finding* intertwined: a point made long ago in Hintikka 1973. In that light, our story so far has only addressed half of the real topic.

9 References

S. Artemov, 1994, 'Logic of Proofs', *Annals of Pure and Applied Logic* 67, 29–59.

S. Artemov, 2005, 'Evidence-Based Common Knowledge', CUNY Graduate Center, New York. Also in *Proceedings TARK 2005*, Singapore.

A. Baltag, H. van Ditmarsch, A. Herzig, T. Hoshi & T. de Lima, 2006, 'The Logic of Iterated Public Announcement', Department of Informatics, IRIT, Toulouse.

A. Baltag, L. Moss and S. Solecki, 1998, 'The Logic of Public Announcements, Common Knowledge and Private Suspicions', *Proceedings TARK 1998*, 43–56. Los Altos: Morgan Kaufmann Publishers. Updated version, department of cognitive science, Indiana University, Bloomington, and department of computing, Oxford University, 2003.

J. van Benthem, 1993, 'Reflections on Epistemic Logic', *Logique et Analyse* 34 (vol. 133-134), 5-14.

J. van Benthem, 2002, 'One is a Lonely Number, on the logic of communication', Tech Report , Institute for Logic, Language and Computation, University o of Amsterdam. In Z. Chatzidakis, P. Koepke & W. Pohlers, eds., 2006, *Logic Colloquium '02,* ASL & A. K. Peters, Wellesley MA, 96 – 129.

J. van Benthem, 2004, 'What One May Come to Know', *Analysis* 64 (282), 95–105.

J. van Benthem, 2006A, 'Epistemic Logic and Epistemology: the state of their affairs', *Philosophical Studies* 128, 2006, 49 - 76.

J. van Benthem, 2006B, 'Logic In Philosophy', Tech Report, ILLC Amsterdam. In D. Jacquette, ed., *Handbook of the Philosophy of Logic*, Elsevier, Amsterdam.

J. van Benthem & E. Pacuit, 2006, 'The Tree of Knowledge in Action', ILLC Amsterdam and *Proceedings AiML 2006*, Melbourne.

P. Blackburn, M. de Rijke and Y. Venema, 2001, *Modal Logic*, Cambridge University Press, Cambridge.

B. Brogaard and J. Salerno, 2002, 'Fitch's Paradox of Knowability', Stanford Electronic Encyclopedia of Philosophy, http://plato.stanford.edu/entries/fitch-paradox/.

H. van Ditmarsch, W. van der Hoek, & B. Kooi, 2006, *Dynamic Epistemic Logic*, to appear with Springer, Dordrecht, Synthese Library.

H. van Ditmarsch & B. Kooi, 2006, 'The Secret of My Success', *Synthese* 151(2), 201 – 232.

D. Edgington, 1985, 'The Paradox of Knowability', *Mind* XCIV, 557 – 568

P. Egré, 2004, *Propositional Attitudes and Epistemic Paradoxes*, Dissertation, University Paris I Panthéon-Sorbonne and IHPST, Paris.

R. Fagin, J. Halpern, Y. Moses & M. Vardi, 1995, *Reasoning about Knowledge*, MIT Press, Cambridge (MA).

D. Gabbay, 1996, *Labelling Deductive Systems (Vol.1),* Clarendon, Oxford.

J. Gerbrandy, 1999, *Bisimulations on Planet Kripke*, Ph.D. thesis, ILLC Amsterdam.

J. Gerbrandy, 2007, 'The Paradox of the Surprise Examination in Dynamic Epistemic Logic', *Synthese* 155:1, 21 – 33.

J. Hintikka, 1973, *Logic, Language Games, and Information*, Oxford University Press, Oxford.

M. Jago, 2006, *Logics for Resource-Bounded Agents*, Ph.D. thesis, School of Computer Science and IT, University of Nottingham.

K. Kelly, 1996, *The Logic of Reliable Inquiry,* Oxford University Press, Oxford.

F. Liu, 2006, 'Diversity of Agents', *Proceedings Workshop on Bounded Rationality*, ESSLLI Summer School. Malaga.

G. E. Moore, 1962, *The Commonplace Book 1919–1953*, Allen & Unwin, London.

M. Osborne & A. Rubinstein, 1994, *A Course in Game Theory*, MIT Press, Cambridge (MA).

J. Sack, 2006, 'Temporal Language for Epistemic Programs', Department of Philosophy, Indiana University, Bloomington.

N. Tennant, 2002, 'Victor Vanquished', *Analysis* 62: 135–142.

A. Yap, 2006, 'Product Update and Looking Backward', ILLC Research Report, University of Amsterdam.