Moorean Phenomena in Epistemic Logic

LORI Workshop, ESSLLI 2010

Wes Holliday and Thomas Icard
Logical Dynamics Lab, CSLI
Department of Philosophy, Stanford University

August 16, 2010

- Under what conditions does a true piece of information remain true when it is received by an agent?

- Under what conditions does a true piece of information remain true when it is received by an agent?
- Sometimes true information becomes false merely in virtue of changes in the external world.

- Under what conditions does a true piece of information remain true when it is received by an agent?

- Sometimes true information becomes false merely in virtue of changes in the external world.

  "It is now 15:02:02"

- Under what conditions does a true piece of information remain true when it is received by an agent?
- Sometimes true information becomes false merely in virtue of changes in the external world.

  "It is now 15:02:02"

- In other cases true information becomes false in virtue of its being received.

- Under what conditions does a true piece of information remain true when it is received by an agent?

- Sometimes true information becomes false merely in virtue of changes in the external world.

  "It is now 15:02:02"

- In other cases true information becomes false in virtue of its being received.

  "You don't know it, but this is the second time ESSLLI has been held in Copenhagen."

- Under what conditions does a true piece of information remain true when it is received by an agent?

- Sometimes true information becomes false merely in virtue of changes in the external world.

  "It is now 15:02:02"

- In other cases true information becomes false in virtue of its being received.

  "You don't know it, but this is the second time ESSLLI has been held in Copenhagen."

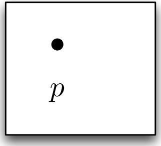
- This talk is about the second kind of case, which is an instance of the *Moore sentence*, of the form $\neg\Box p \wedge p$.

- The Moore sentence $\neg\Box p \wedge p$ is *unsuccessful* in that it does not always remain true when it is learned.

- The Moore sentence $\neg\Box p \wedge p$ is *unsuccessful* in that it does not always remain true when it is learned. In fact, it is *self-refuting* in that it always becomes false when it is learned.

- The Moore sentence $\neg\Box p \wedge p$ is *unsuccessful* in that it does not always remain true when it is learned. In fact, it is *self-refuting* in that it always becomes false when it is learned.

- The Moore sentence $\neg\Box p \wedge p$ is *unsuccessful* in that it does not always remain true when it is learned. In fact, it is *self-refuting* in that it always becomes false when it is learned.



The Moore sentence and Fitch's paradox

- The Moore sentence is also closely tied to Fitch's "paradox of knowability": If there is an unknown truth, then there is an unknowable truth.

The Moore sentence and Fitch's paradox

- The Moore sentence is also closely tied to Fitch's "paradox of knowability": If there is an unknown truth, then there is an unknowable truth.
- If p is true but unknown, then $p \wedge \neg\Box p$ is true. But this latter sentence cannot be known since $\Box(p \wedge \neg\Box p)$ is unsatisfiable, given certain assumptions about knowledge.

Definition

- A formula φ is *successful* just in case for all pointed models $\mathcal{M}, w$, if $\mathcal{M}, w \vDash \varphi$ then $\mathcal{M}_{|\varphi}, w \vDash \varphi$.

Definition

- A formula φ is *successful* just in case for all pointed models $\mathcal{M}, w$, if $\mathcal{M}, w \vDash \varphi$ then $\mathcal{M}_{|\varphi}, w \vDash \varphi$.

In Public Announcement Logic (PAL), we say φ is successful just if, $\vDash [!\varphi]\varphi$.

Definition

- A formula φ is *successful* just in case for all pointed models $\mathcal{M}, w$, if $\mathcal{M}, w \vDash \varphi$ then $\mathcal{M}_{|\varphi}, w \vDash \varphi$.

  In Public Announcement Logic (PAL), we say φ is successful just if, $\vDash [!\varphi]\varphi$.

- A formula φ is *self-refuting* just in case for all pointed models $\mathcal{M}, w$, if $\mathcal{M}, w \vDash \varphi$ then $\mathcal{M}_{|\varphi}, w \nvDash \varphi$.

Definition

- A formula φ is *successful* just in case for all pointed models $\mathcal{M}, w$, if $\mathcal{M}, w \vDash \varphi$ then $\mathcal{M}_{|\varphi}, w \vDash \varphi$.

  In Public Announcement Logic (PAL), we say φ is successful just if, $\vDash [!\varphi]\varphi$.

- A formula φ is *self-refuting* just in case for all pointed models $\mathcal{M}, w$, if $\mathcal{M}, w \vDash \varphi$ then $\mathcal{M}_{|\varphi}, w \nvDash \varphi$.

  In PAL, we say φ is self-refuting just if, $\vDash [!\varphi]\neg\varphi$.

- These formulas have been shown to be at the source of well-known epistemic puzzles in [van Ditmarsch and Kooi, 2006]. See also [Baltag et al., 2008] and [van Benthem, 2004].

- These formulas have been shown to be at the source of well-known epistemic puzzles in [van Ditmarsch and Kooi, 2006]. See also [Baltag et al., 2008] and [van Benthem, 2004].

- A well-known open problem is to give a *syntactic characterization* of the class of the successful formulas, as well as the class of self-refuting formulas.

- These formulas have been shown to be at the source of well-known epistemic puzzles in [van Ditmarsch and Kooi, 2006]. See also [Baltag et al., 2008] and [van Benthem, 2004].

- A well-known open problem is to give a *syntactic characterization* of the class of the successful formulas, as well as the class of self-refuting formulas.

  Is $\neg(p \vee q) \vee (p \wedge (\Box p \vee \Diamond q))$ unsuccessful? Self-refuting?

- In general, it is not easy to tell whether an arbitrary formula is successful or self-refuting.

- In general, it is not easy to tell whether an arbitrary formula is successful or self-refuting.

- The class of successful formulas is not closed under:

- In general, it is not easy to tell whether an arbitrary formula is successful or self-refuting.
- The class of successful formulas is not closed under:
  - Conjunction: $p \wedge \neg \Box p$;

- In general, it is not easy to tell whether an arbitrary formula is successful or self-refuting.
- The class of successful formulas is not closed under:
  - Conjunction: $p \wedge \neg \Box p$;
  - Negation: $\neg(\neg p \vee \Box p)$;

- In general, it is not easy to tell whether an arbitrary formula is successful or self-refuting.
- The class of successful formulas is not closed under:
  - Conjunction: $p \wedge \neg \Box p$;
  - Negation: $\neg(\neg p \vee \Box p)$;
  - Implication: $(\neg p \vee \Box p) \rightarrow \bot$;

- In general, it is not easy to tell whether an arbitrary formula is successful or self-refuting.
- The class of successful formulas is not closed under:
  - Conjunction: $p \wedge \neg \Box p$;
  - Negation: $\neg(\neg p \vee \Box p)$;
  - Implication: $(\neg p \vee \Box p) \rightarrow \bot$;
  - Disjunction: (stay tuned).

Moreover, the validity problem for **S5** can be reduced to the "success problem" and to the "self-refuting problem".

Moreover, the validity problem for **S5** can be reduced to the "success problem" and to the "self-refuting problem". The following result is due to Johan van Benthem.

Theorem

- *The success problem is coNP-complete.*
- *The self-refuting problem is coNP-complete.*

Moreover, the validity problem for **S5** can be reduced to the "success problem" and to the "self-refuting problem". The following result is due to Johan van Benthem.

Theorem

- *The success problem is coNP-complete.*
- *The self-refuting problem is coNP-complete.*

We might conclude there can be no *very simple* characterization.

- The work we describe here is based on a forthcoming paper [Holliday and Icard, 2010] in *Advances in Modal Logic*.

- The work we describe here is based on a forthcoming paper [Holliday and Icard, 2010] in *Advances in Modal Logic*.

- We show that in logics of knowledge and belief for a single agent (extended by **S5**), Moorean phenomena are the source of all self-refutation.

- The work we describe here is based on a forthcoming paper [Holliday and Icard, 2010] in *Advances in Modal Logic*.
- We show that in logics of knowledge and belief for a single agent (extended by **S5**), Moorean phenomena are the source of all self-refutation.
- Moreover, in logics for an introspective agent (extending **KD45**), Moorean phenomena are the source of all unsuccessfulness as well.

- The work we describe here is based on a forthcoming paper [Holliday and Icard, 2010] in *Advances in Modal Logic*.
- We show that in logics of knowledge and belief for a single agent (extended by **S5**), Moorean phenomena are the source of all self-refutation.
- Moreover, in logics for an introspective agent (extending **KD45**), Moorean phenomena are the source of all unsuccessfulness as well.
- Syntactic characterizations of the two classes of formulas are also obtained in an appendix. They are somewhat complicated.

- First we give definitions codifying the notions of a *Moore sentence*, and a *Moorean sentence*.

- First we give definitions codifying the notions of a *Moore sentence*, and a *Moorean sentence*. For example:
  - $p \wedge \neg\Box p$ is a *Moore sentence*;

- First we give definitions codifying the notions of a *Moore sentence*, and a *Moorean sentence*. For example:
  - $p \wedge \neg \Box p$ is a *Moore sentence*; so is $p \wedge \neg \Box q \wedge \Box (p \rightarrow q)$.

- First we give definitions codifying the notions of a *Moore sentence*, and a *Moorean sentence*. For example:

  - $p \wedge \neg\Box p$ is a *Moore sentence*; so is $p \wedge \neg\Box q \wedge \Box(p \rightarrow q)$.
  - $p \wedge \neg\Box q$ is a *Moorean sentence*;

- First we give definitions codifying the notions of a *Moore sentence*, and a *Moorean sentence*. For example:
  - $p \wedge \neg\Box p$ is a *Moore sentence*; so is $p \wedge \neg\Box q \wedge \Box(p \rightarrow q)$.
  - $p \wedge \neg\Box q$ is a *Moorean sentence*; but $p \wedge \neg\Box q \wedge \neg\Box(p \rightarrow q)$ is not.

- First we give definitions codifying the notions of a *Moore sentence*, and a *Moorean sentence*. For example:
  - $p \wedge \neg\Box p$ is a *Moore sentence*; so is $p \wedge \neg\Box q \wedge \Box(p \rightarrow q)$.
  - $p \wedge \neg\Box q$ is a *Moorean sentence*; but $p \wedge \neg\Box q \wedge \neg\Box(p \rightarrow q)$ is not.

Theorem

- *If a formula is self-refuting in any sublogic of* **S5***, then it is a Moore sentence.*
- *If a formula is unsuccessful in any extension of* **KD45***, then it is a Moorean sentence.*

- First we give definitions codifying the notions of a *Moore sentence*, and a *Moorean sentence*. For example:
  - $p \wedge \neg\Box p$ is a *Moore sentence*; so is $p \wedge \neg\Box q \wedge \Box(p \rightarrow q)$.
  - $p \wedge \neg\Box q$ is a *Moorean sentence*; but $p \wedge \neg\Box q \wedge \neg\Box(p \rightarrow q)$ is not.

Theorem

- *If a formula is self-refuting in any sublogic of* **S5**, *then it is a Moore sentence.*
- *If a formula is unsuccessful in any extension of* **KD45**, *then it is a Moorean sentence.*

Notably, the converses do not hold in general. Moreover, the underlying logics are crucial.

What are the sources of unsuccessfulness in logics for an agent without introspection (logics without axioms **4** and **5**)?

What are the sources of unsuccessfulness in logics for an agent without introspection (logics without axioms **4** and **5**)?

- From an epistemic perspective, the most interesting (normal) proper sublogics of **S5** are obtained by dropping axiom **5** and adding something weaker in its place.

What are the sources of unsuccessfulness in logics for an agent without introspection (logics without axioms **4** and **5**)?

- From an epistemic perspective, the most interesting (normal) proper sublogics of **S5** are obtained by dropping axiom **5** and adding something weaker in its place.
- Logics such as **S4**, **S4.x** for **x**=**2**,**3**,**4**, etc., have been proposed as logics of knowledge.

What are the sources of unsuccessfulness in logics for an agent without introspection (logics without axioms **4** and **5**)?

- From an epistemic perspective, the most interesting (normal) proper sublogics of **S5** are obtained by dropping axiom **5** and adding something weaker in its place.
- Logics such as **S4**, **S4.x** for **x**=**2**,**3**,**4**, etc., have been proposed as logics of knowledge.

Call logics **L** and **L**′ *comparable* if **L** is a sublogic of **L**′ or *vice versa*.

What are the sources of unsuccessfulness in logics for an agent without introspection (logics without axioms **4** and **5**)?

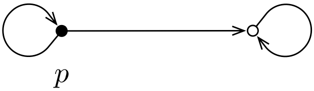
- From an epistemic perspective, the most interesting (normal) proper sublogics of **S5** are obtained by dropping axiom **5** and adding something weaker in its place.
- Logics such as **S4**, **S4.x** for **x**=**2**,**3**,**4**, etc., have been proposed as logics of knowledge.

Call logics **L** and **L**′ *comparable* if **L** is a sublogic of **L**′ or *vice versa*.

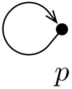
Proposition

For any normal, proper sublogic **L** *of* **S5**, *comparable to* **S4.4**, *there is a formula* (*consistent with* **S5**) *that is unsuccessful in* **L** *but is not Moorean.*

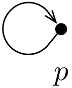
Consider $\Diamond p \wedge \Diamond \neg p$ and the **S4.4** model in the figure below.



Consider $\Diamond p \wedge \Diamond \neg p$ and the **S4.4** model in the figure below.



Consider $\Diamond p \wedge \Diamond \neg p$ and the **S4.4** model in the figure below.



The example shows that without negative introspection, one can come to know p by being truly told, "You do not know whether or not p."

Proposition

For any normal, proper sublogic **L** *of* **S5***, comparable to* **S4.4***, there is a formula (consistent with* **S5***) that is unsuccessful in* **L** *but is not Moorean.*

S5 is unique among the typical logics of knowledge and **KD45** unique among typical logics of belief, insofar as all of their unsuccessful formulas are Moorean.

Proposition

For any normal, proper sublogic **L** *of* **S5**, *comparable to* **S4.4**, *there is a formula (consistent with* **S5**) *that is unsuccessful in* **L** *but is not Moorean.*

S5 is unique among the typical logics of knowledge and **KD45** unique among typical logics of belief, insofar as all of their unsuccessful formulas are Moorean.

For agents without introspection, there are non-Moorean sources of unsuccessfulness.

Theorem

- *If a formula is self-refuting in any sublogic of* **S5**, *then it is a Moore sentence.*
- *If a formula is unsuccessful in any extension of* **KD45**, *then it is a Moorean sentence.*

Theorem

- *If a formula is self-refuting in any sublogic of* **S5**, *then it is a Moore sentence.*
- *If a formula is unsuccessful in any extension of* **KD45**, *then it is a Moorean sentence.*

Neither the converse of (i) nor the converse of (ii) holds in general. Understanding why the converses fail leads to interesting connections with other formula classes.

How can a Moore sentence fail to be self-refuting?

How can a Moore sentence fail to be self-refuting?

Definition

- A formula φ is (*potentially*) *informative* iff there is a pointed model such that $\mathcal{M}, w \vDash \varphi$ and $\mathcal{M}_{|\varphi} \neq \mathcal{M}$. Otherwise φ is *uninformative*.

How can a Moore sentence fail to be self-refuting?

Definition

- A formula φ is (*potentially*) *informative* iff there is a pointed model such that $\mathcal{M}, w \vDash \varphi$ and $\mathcal{M}_{|\varphi} \neq \mathcal{M}$. Otherwise φ is *uninformative*.
- A formula φ is *always informative* iff for all pointed models such that $\mathcal{M}, w \vDash \varphi$, $\mathcal{M}_{|\varphi} \neq \mathcal{M}$.

How can a Moore sentence fail to be self-refuting?

Definition

- A formula φ is (*potentially*) *informative* iff there is a pointed model such that $\mathcal{M}, w \vDash \varphi$ and $\mathcal{M}_{|\varphi} \neq \mathcal{M}$. Otherwise φ is *uninformative*.
- A formula φ is *always informative* iff for all pointed models such that $\mathcal{M}, w \vDash \varphi$, $\mathcal{M}_{|\varphi} \neq \mathcal{M}$.

If a formula is not always informative, then it is not self-refuting, for there is a model such that $\mathcal{M}, w \vDash \varphi$ but $\mathcal{M}_{|\varphi} = \mathcal{M}$, so $\mathcal{M}_{|\varphi}, w \vDash \varphi$.

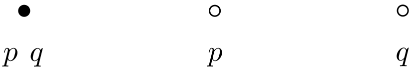
Perhaps a sentence is self-refuting iff it is a Moore sentence *and* it is always informative?

Perhaps a sentence is self-refuting iff it is a Moore sentence *and* it is always informative?

Example: $(p \wedge \Diamond\neg p) \vee (p \wedge q \wedge \Diamond\neg q)$ is always informative but not self-refuting.

Perhaps a sentence is self-refuting iff it is a Moore sentence *and* it is always informative?

Example: $(p \wedge \Diamond\neg p) \vee (p \wedge q \wedge \Diamond\neg q)$ is always informative but not self-refuting.



Perhaps a sentence is self-refuting iff it is a Moore sentence *and* it is always informative?

Example: $(p \wedge \Diamond\neg p) \vee (p \wedge q \wedge \Diamond\neg q)$ is always informative but not self-refuting.

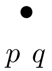
Perhaps a sentence is self-refuting iff it is a Moore sentence *and* it is always informative?

Example: $(p \wedge \Diamond\neg p) \vee (p \wedge q \wedge \Diamond\neg q)$ is always informative but not self-refuting.

However, the formula is *self-refuting within two steps*.

Perhaps a sentence is self-refuting iff it is a Moore sentence *and* it is always informative?

Example: $(p \wedge \Diamond \neg p) \vee (p \wedge q \wedge \Diamond \neg q)$ is always informative but not self-refuting.



However, the formula is *self-refuting within two steps*.

Related Classes of Formulas

Perhaps a sentence is self-refuting iff it is a Moore sentence *and* it is always informative?

Example: $(p \wedge \Diamond \neg p) \vee (p \wedge q \wedge \Diamond \neg q)$ is always informative but not self-refuting.

However, the formula is *self-refuting within two steps*. This example points to the interest of self-refutation "in the long run."

Definition

Given a model $\mathcal{M}$, we define $\mathcal{M}_{|^n\varphi}$ recursively by $\mathcal{M}_{|^0\varphi} = \mathcal{M}$, $\mathcal{M}_{|^{n+1}\varphi} = \left(\mathcal{M}_{|^n\varphi}\right)_{|\varphi}$. A formula φ is *eventually self-refuting* iff for all pointed models, if $\mathcal{M}, w \vDash \varphi$, then there is an n such that $\mathcal{M}_{|^n\varphi}, w \nvDash \varphi$.

Definition

Given a model $\mathcal{M}$, we define $\mathcal{M}_{|^n\varphi}$ recursively by $\mathcal{M}_{|^0\varphi} = \mathcal{M}$, $\mathcal{M}_{|^{n+1}\varphi} = \left(\mathcal{M}_{|^n\varphi}\right)_{|\varphi}$. A formula φ is *eventually self-refuting* iff for all pointed models, if $\mathcal{M}, w \vDash \varphi$, then there is an n such that $\mathcal{M}_{|^n\varphi}, w \nvDash \varphi$.

One more definition, not obviously related:

Definition

φ is *Cartesian* iff $\Box\varphi$ is satisfiable.

Definition

Given a model $\mathcal{M}$, we define $\mathcal{M}_{|^n\varphi}$ recursively by $\mathcal{M}_{|^0\varphi} = \mathcal{M}$, $\mathcal{M}_{|^{n+1}\varphi} = \left(\mathcal{M}_{|^n\varphi}\right)_{|\varphi}$. A formula φ is *eventually self-refuting* iff for all pointed models, if $\mathcal{M}, w \vDash \varphi$, then there is an n such that $\mathcal{M}_{|^n\varphi}, w \nvDash \varphi$.

One more definition, not obviously related:

Definition

φ is *Cartesian* iff $\Box\varphi$ is satisfiable.

Proposition

The following are equivalent:

1. φ *is always informative.*
2. φ *is not Cartesian.*
3. φ *is eventually self-refuting.*

Proposition

The following are equivalent:

1. φ *is always informative.*
2. φ *is not Cartesian.*
3. φ *is eventually self-refuting.*

In other words, **the sentences that always provide information to an agent, no matter the agent's prior epistemic state, are exactly those sentences that cannot be known—and will eventually become false if repeated enough**.

Theorem

- *If a formula is self-refuting in any sublogic of* **S5***, then it is a Moore sentence.*
- *If a formula is unsuccessful in any extension of* **KD45***, then it is a Moorean sentence.*

Neither the converse of (i) nor the converse of (ii) holds in general. Understanding why the converses fail leads to other interesting results.

How can a Moorean sentence fail to be unsuccessful?

How can a Moorean sentence fail to be unsuccessful?

Example: $(p \wedge \Diamond\neg p) \vee \Box p$ and $(p \wedge \Diamond q) \vee \Box p$ are both Moorean sentences according to our definition, but they are both successful. The reason is a kind of *compensation*.

How can a Moorean sentence fail to be unsuccessful?

Example: $(p \wedge \Diamond\neg p) \vee \Box p$ and $(p \wedge \Diamond q) \vee \Box p$ are both Moorean sentences according to our definition, but they are both successful. The reason is a kind of *compensation*.

However, you can only compensate so much...

Definition

A formula φ is *super-successful* iff for every pointed model, $\mathcal{M}, w \vDash \varphi$ implies $\mathcal{M}', w \vDash \varphi$ for every $\mathcal{M}'$ such that $\mathcal{M}_{|\varphi} \subseteq \mathcal{M}' \subseteq \mathcal{M}$.

Definition

A formula φ is *super-successful* iff for every pointed model, $\mathcal{M}, w \vDash \varphi$ implies $\mathcal{M}', w \vDash \varphi$ for every $\mathcal{M}'$ such that $\mathcal{M}_{|\varphi} \subseteq \mathcal{M}' \subseteq \mathcal{M}$.

- If φ is super-successful and $\mathcal{M}, w \vDash \varphi$, then as points that are not in $\mathcal{M}_{|\varphi}$ are eliminated from $\mathcal{M}$, φ remains true at w.

Definition

A formula φ is *super-successful* iff for every pointed model, $\mathcal{M}, w \vDash \varphi$ implies $\mathcal{M}', w \vDash \varphi$ for every $\mathcal{M}'$ such that $\mathcal{M}_{|\varphi} \subseteq \mathcal{M}' \subseteq \mathcal{M}$.

- If φ is super-successful and $\mathcal{M}, w \vDash \varphi$, then as points that are not in $\mathcal{M}_{|\varphi}$ are eliminated from $\mathcal{M}$, φ remains true at w.
- Since we take the elimination of points as an agent's acquisition of new information, this means that φ remains true as the agent approaches, by way of the incremental acquisition of new information, the epistemic state of $\mathcal{M}_{|\varphi}$ wherein the agent knows φ.

Definition

A formula φ is *super-successful* iff for every pointed model, $\mathcal{M}, w \vDash \varphi$ implies $\mathcal{M}', w \vDash \varphi$ for every $\mathcal{M}'$ such that $\mathcal{M}_{|\varphi} \subseteq \mathcal{M}' \subseteq \mathcal{M}$.

- If φ is super-successful and $\mathcal{M}, w \vDash \varphi$, then as points that are not in $\mathcal{M}_{|\varphi}$ are eliminated from $\mathcal{M}$, φ remains true at w.
- Since we take the elimination of points as an agent's acquisition of new information, this means that φ remains true as the agent approaches, by way of the incremental acquisition of new information, the epistemic state of $\mathcal{M}_{|\varphi}$ wherein the agent knows φ.
- Intuitively, we can say that a super-successful formula remains true while an agent is "on the way" to learning it.

Proposition

Not all successful formulas are super-successful.

Proposition

Not all successful formulas are super-successful.

In other words, **there are sentences that always remain true when they are learned, but whose truth value may oscillate while an agent is on the way to learning them**.

Proposition

Not all successful formulas are super-successful.

This proposition has several interesting corollaries, together with the following.

Proposition

Not all successful formulas are super-successful.

This proposition has several interesting corollaries, together with the following.

Proposition

If φ is not super-successful, then there is a successful formula ψ such that $\varphi \vee \psi$ is unsuccessful.

Proposition

Not all successful formulas are super-successful.

This proposition has several interesting corollaries, together with the following.

Proposition

If φ is not super-successful, then there is a successful formula ψ such that $\varphi \vee \psi$ is unsuccessful.

A surprising failure of closure is immediate from the previous propositions.

Corollary

The set of successful formulas is not closed under disjunction.

From the previous result, we can draw a connection with the *learnable* (a.k.a. *knowable*) formulas, introduced in [van Benthem, 2004].

From the previous result, we can draw a connection with the *learnable* (a.k.a. *knowable*) formulas, introduced in [van Benthem, 2004].

Definition

A formula φ is (*always*) *learnable* iff for all pointed models, if $\mathcal{M}, w \vDash \varphi$, then there is some ψ such that $\mathcal{M}_{|\psi}, w \vDash \Box\varphi$.

As noted in [Balbiani et al., 2008], all successful formulas are learnable.

From the previous result, we can draw a connection with the *learnable* (a.k.a. *knowable*) formulas, introduced in [van Benthem, 2004].

Definition

A formula φ is (*always*) *learnable* iff for all pointed models, if $\mathcal{M}, w \vDash \varphi$, then there is some ψ such that $\mathcal{M}_{|\psi}, w \vDash \Box\varphi$.

As noted in [Balbiani et al., 2008], all successful formulas are learnable.

However, the following is immediate from the fact that successful formulas are not closed under disjunction.

Corollary

Not all learnable formulas are successful.

Corollary

Not all learnable formulas are successful.

In other words, **there are sentences that sometimes become false when learned directly, but which an agent can always come to know indirectly by learning something else**.

We have seen why not every Moore sentence is self-refuting and why not every Moorean sentence is unsuccessful.

We have seen why not every Moore sentence is self-refuting and why not every Moorean sentence is unsuccessful.

This lead to interesting results relating self-refuting and unsuccessful formulas to other formula classes: *always informative*, *Cartesian*, *eventually self-refuting*, *super-successful*, and *learnable* formulas.

We have seen why not every Moore sentence is self-refuting and why not every Moorean sentence is unsuccessful.

This lead to interesting results relating self-refuting and unsuccessful formulas to other formula classes: *always informative*, *Cartesian*, *eventually self-refuting*, *super-successful*, and *learnable* formulas.

But what about a full characterization of self-refuting and (un)successful formulas?

We have seen why not every Moore sentence is self-refuting and why not every Moorean sentence is unsuccessful.

This lead to interesting results relating self-refuting and unsuccessful formulas to other formula classes: *always informative*, *Cartesian*, *eventually self-refuting*, *super-successful*, and *learnable* formulas.

But what about a full characterization of self-refuting and (un)successful formulas?

Theorem

1. *A formula is self-refuting iff it is a* strong *Moore sentence.*
2. *A formula is unsuccessful iff it is a* strong *Moorean sentence.*

Review of some main points:

- For introspective agents, the only true sentences that may become false when learned are variants of the Moore sentence.

Review of some main points:

- For introspective agents, the only true sentences that may become false when learned are variants of the Moore sentence.
- For agents without introspection, there are non-Moorean sources of unsuccessfulness.

- The sentences that always provide information to an agent, no matter the agent's prior epistemic state, are exactly those sentences that cannot be known—and will eventually become false if repeated enough.

- The sentences that always provide information to an agent, no matter the agent's prior epistemic state, are exactly those sentences that cannot be known—and will eventually become false if repeated enough.
- There are sentences that always remain true when they are learned, but whose truth value may oscillate while an agent is on the way to learning them.

- The sentences that always provide information to an agent, no matter the agent's prior epistemic state, are exactly those sentences that cannot be known—and will eventually become false if repeated enough.
- There are sentences that always remain true when they are learned, but whose truth value may oscillate while an agent is on the way to learning them.
- There are sentences that sometimes become false when learned directly, but which an agent can always come to know indirectly by learning something else.

- The formulas that are self-refuting are exactly the *strong* Moore sentences, and the formula that are unsuccessful are exactly the *strong* Moorean sentences.

Thank you!

Balbiani, P., Baltag, A., van Ditmarsch, H., Herzig, A., Hoshi, T., and de Lima, T. (2008).
'Knowable' as 'known after an announcement'.
The Review of Symbolic Logic, 1:305–334.

Baltag, A., van Ditmarsch, H., and Moss, L. (2008).
Epistemic logic and information update.
In Adriaans, P. and van Benthem, J., editors, *Philosophy of Information*, pages 361–456. North Holland.

van Benthem, J. (2004).
What one may come to know.
Analysis, 64(2):95–105.

van Ditmarsch, H. and Kooi, B. (2006).
The secret of my success.
Synthese, 151:201–232.

Holliday, W. and Icard, T. (2010).
Moorean phenomena in epistemic logic.
In Beklemishev, L., Goranko, V., and Shehtman, V., editors, *Advances in Modal Logic*, pages 168–187. College Publications.