# Truth in Fiction via Non-Standard Belief Revision

Christopher Badura
Ruhr-Universität Bochum

CL16, 12th September 2016

# Outline

Truth in Fiction
- Fiction operator "$In_f$"
- Lewis's Analysis 2 for $In_f$
- A Problem with inconsistency

Finding Semantics for $In_f$
- Plausibility Models
- Belief Revision - Soft Upgrades
- Solving the Problem

Conclusion

# The Problem of Truth in Fiction

Anyone who utters:

(S) Sherlock Holmes lives in Baker Street

would not be objected against by non-philosophers.

However:

- What if Sherlock Holmes does not exist?
- What if Sherlock Holmes exists?
- (S) clashes with known fact.

# Solution: Fiction-Operator

Claim: (S) is elliptical for '*In the fiction f*, Sherlock Holmes lives in Baker Street', where $f$ is the relevant fiction.

$\Rightarrow$ Introduce a sentential operator $In_f$, where '$f$' is a variable for the name of the relevant fiction.

So, literally, (S) is false, but it is true that $In_f, (S)$

Question: What are the truth conditions for $In_f$, what is its semantics?

$In_f$ is an (hyper-)intensional operator
$\Rightarrow$ (im)possible world semantics

# Desiderata for an Analysis

The desired semantics should account for

1. Explicit Content
2. Import of background knowledge/beliefs
3. Logical consequences
4. Inconsistent fictions without trivializing

# Lewis's Analysis 2

Paraphrasing David Lewis (1978):

(Analysis 2) A sentence of the form 'In the fiction $f$, $\varphi$' is non-vacuously true iff. for every collective belief world $w$ of the community of origin of $f$ there is a world $v$ such that

1. $f$ is told as known fact in $v$
2. $\varphi$ is true at $v$
3. $v$ differs less from the world $w$, on balance, than does any world $u$ where $f$ is told as known fact and $\varphi$ is not true.

It is vacuously true iff. there are no possible worlds where $f$ is told as known fact.

# Desiderata Check

1. ✓ by $f$ is told as known fact
2. ✓ by considering collective belief worlds
3. ✓ by considering possible worlds, closed under classical logic
4. ✗ if a fiction is inconsistent, there is no world where it is told as known fact, thus everything is true in it and thus all inconsistent fictions have the same content

# The Object Language



Our object language is a propositional language where $Prop = \{p, q, r, ...\}$ is a countable set of propositional variables or atomic sentences. We generate the well formed formulas by

$$p | \neg\varphi | \varphi \wedge \psi | \varphi \vee \psi | \varphi \supset \psi | \diamond \varphi | \Box\varphi | In_f, \varphi$$

where $p \in Prop$

A *Single-Agent Plausibility Model* is a tuple $\mathcal{M} = (W, \leq, V)$, where

- $W = P \cup I$ is a non-empty set of possible worlds $P$ and a set of impossible worlds $I$, s.t. $P \cap I = \emptyset$
- $\leq \subseteq W \times W$ is the agent's plausibility order. Transitive conversely well-founded relation. Not necessarily anti-symmetric. Converse well-foundedness implies reflexivity and totality
  $w \simeq v$ iff. $w \leq v$ and $v \leq w$
  $w < v$ iff. $w \leq v$ and $v \not\leq w$ (not $v \leq w$)
- $V$ is a pair $(V^+, V^-)$ s.t. $V^{\pm} : Prop \to \mathcal{P}(W)$ and $V^+(p)$ is the set of worlds where $p$ is true and $V^-(p)$ is the set of worlds where $p$ is false.

# Semantics

## Truth

If $w \in P$, we define truth/falsity at a world in a model recursively:

$$w \vDash^+ \varphi \text{ iff. } w \in V^+(\varphi), \text{ for } \varphi \in Prop$$
$$w \vDash^- \varphi \text{ iff. } w \in V^-(\varphi), \text{ for } \varphi \in Prop$$

$$w \vDash^+ \neg\varphi \text{ iff. } w \vDash^- \varphi$$
$$w \vDash^- \neg\varphi \text{ iff. } w \vDash^+ \varphi$$

$$w \vDash^+ \varphi \wedge \psi \text{ iff. } w \vDash^+ \varphi \text{ and } w \vDash^+ \psi$$
$$w \vDash^- \varphi \wedge \psi \text{ iff. } w \vDash^- \varphi \text{ or } w \vDash^- \psi$$

$$w \vDash^+ \varphi \vee \psi \text{ iff. } w \vDash^+ \varphi \text{ or } w \vDash^+ \psi$$
$$w \vDash^- \varphi \vee \psi \text{ iff. } w \vDash^- \varphi \text{ and } w \vDash^- \psi$$

$$w \vDash^+ \varphi \supset \psi \text{ iff. } w \vDash^+ \varphi \text{ implies } w \vDash^+ \psi$$
$$w \vDash^- \varphi \supset \psi \text{ iff. } w \vDash^+ \varphi \text{ and } w \vDash^- \psi$$

Note that implication is material implication.

# Semantics

## Truth-Continued

And the modalities:

$w \models^{+} \Box\varphi$ iff. for all $v \in P$: $v \models^{+} \varphi$
$w \models^{-} \Box\varphi$ iff. for some $v \in P$: $v \models^{-} \varphi$

$w \models^{+} \Diamond\varphi$ iff. for some $v \in P$: $v \models^{+} \varphi$
$w \models^{-} \Diamond\varphi$ iff. for all $v \in P$: $v \models^{-} \varphi$

$w \models^{+} In_f, \varphi$ iff ???
$w \models^{-} In_f, \varphi$ iff. ???



Christopher Badura
Ruhr-Universität Bochum

# Classicality Condition

We impose a classicality condition (CC) on the possible worlds, that is

(CC) For all $w \in P$ and all $\varphi \in Form$: either $w \vDash^{+} \varphi$ or $w \vDash^{-} \varphi$ and not both.

# Semantics

## Truth-Continued

For impossible worlds, we extend $V^+$ and $V^-$ to arbitrary formulas, so they assign sets of impossible worlds to formulas in a direct manner. So if $w \in I$, then

$$w \models^+ \varphi \text{ iff. } w \in V^+(\varphi)$$
$$w \models^- \varphi \text{ iff. } w \in V^-(\varphi)$$

# Semantics

## Logical Consequence

Logical Consequence is defined over the set of possible worlds $P$. Let $\Gamma$ be a set of formulas and $\varphi$ a formula. Then

$$\Gamma \vDash \varphi \text{ if for any model } \langle W, \leq, V \rangle \text{ and any } w \in P : \\ \text{if } w \vDash^{+} \gamma \text{ for any } \gamma \in \Gamma, \text{ then } w \vDash^{+} \varphi$$

We then say $\varphi$ is a *logical consequence* of $\Gamma$. $\varphi$ is a *logical truth* iff. $\vDash \varphi$ iff. $\emptyset \vDash \varphi$.

# Soft Upgrades

Let $F$ be the explicit content of a fiction $f$. We treat engagement with fiction as a soft upgrade expressed by $\Uparrow F$. That is, all worlds where $F$ is true ($f$-worlds) become more plausible than any worlds where $F$ is not true (non-$f$-worlds).
We write $w \models^{+} F$ if for all $\varphi \in F$, $w \models^{+} \varphi$.
We impose two conditions on the new ordering $\leq^{\Uparrow F}$:

Cond 1: $\forall t \in W[t \vDash^{+} F \Rightarrow (\forall s \in W(s \nvDash^{+} F \Rightarrow s \leq^{\Uparrow F} t))]$

To ensure that the most plausible worlds in engagement with fiction are $f$-worlds, we have:

Cond 2: $\forall t \in W[(\forall s \in W : s \leq^{\Uparrow F} t) \Rightarrow t \vDash^{+} F]$

Among the $f$-worlds, and also among the non-$f$-worlds, we assume the ordering to be given externally by the agent. It can remain the same or change arbitrarily but has to remain a transitive conversely well-founded order.

The single agent models can be extended to multi-agent models by having a set of agents $\mathcal{A}$ and a set of plausibility orderings $\{\leq_a\}_{a\in\mathcal{A}}$.
The crucial part is to get a group ordering $\leq_{\mathcal{G}}^{\Uparrow F}$ on the worlds. Our way is based on a hierarchy among agents, but one could also use ideas from social choice theory.

# Group ordering

hierarchy among agents $a$ in some group $\mathcal{G} \subseteq \mathcal{A}$ with $n$ agents. Based on the hierarchy $a_0, \ldots, a_n$, where $a_0$ is the highest in the hierarchy, define the group ordering $\leq^{\Uparrow F}_{\mathcal{G}:n}$ inductively:

$$\leq^{\Uparrow F}_{\mathcal{G}_0/a_1} = <^{\Uparrow F}_{a_0} \cup (\simeq^{\Uparrow F}_{a_0} \cap \leq^{\Uparrow F}_{a_1})$$

$$\leq^{\Uparrow F}_{\mathcal{G}_n/a_{n+1}} = <^{\Uparrow F}_{\mathcal{G}_n} \cup (\simeq^{\Uparrow F}_{\mathcal{G}_n} \cap \leq^{\Uparrow F}_{a_{n+1}})$$

If for every agent $a$, $\leq^{\Uparrow F}_{a}$ is a transitive conversely well-founded relation satisfying Cond 1 and Cond 2, then $\leq^{\Uparrow F}_{\mathcal{G}}$ satisfies those conditions too.

# Overt/Common Beliefs





Instead of Lewis's collective belief worlds we will use common belief worlds.

On our models, we can define a common belief modality $CB_{\mathcal{G}}$. $\varphi$ is commonly believed if everyone believes $\varphi$ and everyone believes that everyone believes $\varphi$ *and so on*.

If $CB_{\mathcal{G}}^{w} = \{\varphi | w \vDash^{+} CB_{\mathcal{G}}\varphi\}$ is the set of common beliefs of group $\mathcal{G}$ at $w$, then we let $|CB_{\mathcal{G}}^{w}|$ be the set of worlds, where all those common beliefs are true.

# Semantics for '$In_f$'



Let $\mathcal{M}$ be a multi-agent plausibility model where $\mathcal{G} \subseteq \mathcal{A}$ is the community of origin of $f$. Let $w \in P$. For $S \subseteq W$, we define $best_{\mathcal{G}}^{\Uparrow F} S = \{w \in W | \forall x \in S : x \leq_{\mathcal{G}}^{\Uparrow F} w\}$. Then

$$\mathcal{M}, w \vDash^{+} In_f, \varphi \text{ iff. } \forall v \in best_{\mathcal{G}}^{\Uparrow F} |CB_{\mathcal{G}}^{w}| (v \vDash^{+} F \Rightarrow v \vDash^{+} \varphi)$$

$$\mathcal{M}, w \vDash^{-} In_f, \varphi \text{ iff. } \exists v \in best_{\mathcal{G}}^{\Uparrow F} |CB_{\mathcal{G}}^{w}| (v \vDash^{+} F \text{ and } v \nvDash^{+} \varphi)$$

# Inconsistent Fictions Again

Does $In_f, (\varphi \wedge \neg\varphi) \vDash In_f, (\psi)$ hold?

No! We can easily have a most plausible impossible world after the upgrade which makes $F$ and $\varphi \wedge \neg\varphi$ true but not $\psi$. The easiest case if $\varphi \wedge \neg\varphi$ is in the explicit content $F$.

# Some other Results

- $In_f, F$ is a logical truth since logical truth is defined only over possible worlds. This is nice since every fiction makes its explicit content true.
- It seems, every logical inference within the scope of the operator can fail.

  Reasonable if one accepts that fiction can be about anything, and thus about any logic failing. Clearly, this depends on the fiction in question and the plausibility orderings.

# Conclusion

## Desiderata Check

1. Explicit content: ✓ because $In_f, F$ is a logical truth
2. Import of background knowledge/beliefs: (✓) depends on the plausibility ordering
3. Import of logical consequences: (✓) depends on the plausibility ordering
4. Inconsistent fiction without trivializing: ✓ the inference from $In_f, (\varphi \wedge \neg\varphi)$ to $In_f, (\psi)$ fails in general.

Starting from Lewis's Analysis 2, we arrived at a semantics for $In_f$ which avoids the trouble with inconsistent fiction.

# Thank You!

christopher.badura@rub.de

# Common Belief modality

Define $s(a) = \{w \in W | w \simeq_a s\}$
Define $s \rightarrow_a t$ iff. $t \in bests(a)$
Define $\rightarrow_{CB}$ as the smallest relation $R$ such that $R \subseteq \bigcup_a \rightarrow_a$
Define $w \models^+ CB_{\mathcal{G}}\varphi$ iff. for all $v$ such that $w \rightarrow_{CB} v$ $v \models^+ \varphi$.