


Thought Experiments and Modal Logic: fixing Sorensen's and Häggqvist's schemas

EPSA 2023

Ruward Mulder, Cambridge University
([Erkenntnis 2023](#), joint work with F.A. Muller, Erasmus University Rotterdam)

September 20th, 2023






I. Sorensen & Häggqvist

II. Searle's Chinese Room

III. (In)consistency

IV. Moving forward



I. Sorensen & Häggqvist

II. Searle's Chinese Room

III. (In)consistency

IV. Moving forward

Destructive thought experiments

The aim is to undermine some theory T by imagining an absurd scenario were T to be true

They often spawn many opposing views: controversy

Destructive TE	Target theory (T)	Academic area
Schrödinger's cat	Quantum mechanics	Physics
Trolley problem	Utilitarianism	Ethics
Chinese Room Argument	Strong A.I.	Computer science
Leibniz kinematical shift	Absolute space	Classical mechanics
Gould's 'Replaying life's tape'	Evolutionary determinism	Evolution/archeology
What Mary didn't know	Physicalism	Mind
Neurath's Scientific Utopianism	Capitalism	Economic/social theory
Maxwell's Demon	Acceptationless 2nd law	Kinetic theory
...

Searching for structure

What kind of structures do such disparate thought experiments have in common?

Focus on: argument structure (akin to the Argument View (Norton 2004))

Need not challenge other views (cf. Mauricio Suárez, forthcoming): those that focus on narrative structure, cognitive aspects, imagination and/or fiction view

(cf. Tamar Gendler, Mike Stuart, Elke Brendel, Sam Rijken, ...)

Heuristic value!

- Abstract level: identifying patterns of disagreements (insights into disparities between modal intuitions)
- Particular debates: **classification of controversy** (accounting for differences between opposing views)

Three modal-logical schemas for destructive TEs

For the usual alethic operators: $\diamond A$ ('it is possible that');
 $\Box A$ ('it is necessary that');
 $A \Box \rightarrow B$ ('counterfactual: if A were the case then B')

Sorensen's Necessity Refuter:

1. T
2. $T \rightarrow \Box I$
3. $(I \& C) \Box \rightarrow W$
4. $\neg \diamond W$
5. $\diamond C$

Sorensen's Possibility Refuter:

- I. T
- II. $T \rightarrow \diamond I$
- III. $(I \& C) \Box \rightarrow W$
- IV. $\neg \diamond W$
- V. $\diamond I \rightarrow \diamond (I \& C)$

Häggqvist's Counterfactual Refuter:

- i. T
- ii. $\diamond C$
- iii. $T \rightarrow (C \Box \rightarrow W)$
- iv. $(C \Box \rightarrow \neg W)$

For each schema the premises are mutually inconsistent:
→ one of them needs to go!

There are then 5 ways (Sorensen) or 4 ways (Häggqvist) to disagree.

Thus, successful application furthers understanding: insights into *why* the disagreeers disagree amongst themselves



I. Sorensen & Häggqvist

II. Searle's Chinese Room

III. (In)consistency

IV. Moving forward

Illustration: Searle's Chinese room thought experiment



The Necessity Refuter and the Chinese room argument

1. T
 2. $T \rightarrow \Box I$
 3. $(I \& C) \Box \rightarrow W$
 4. $\neg \Diamond W$
 5. $\Diamond C$
1. The theory of strong A.I.
 2. According to strong A.I., the implementation of the algorithm passes the Turing test
 3. Were Searle to hand-implement the algorithm and pass the Turing test, then he would understand Chinese
 4. But that's absurd! Searle says he does not understand Chinese
 5. It is possible that Searle hand-implements this algorithm

The classification of controversy

Sorensen's Necessity Refuter

1. T
2. $T \rightarrow \Box I$
3. $(I \& C) \Box \rightarrow W$
4. $\neg \Diamond W$
5. $\Diamond C$

Häggqvist's Counterfactual Refuter

- i. T
- ii. $\Diamond C$
- iii. $T \rightarrow (C \Box \rightarrow W)$
- iv. $(C \Box \rightarrow \neg W)$

Replies to CRA	Sorensen's Necessity Refuter	Häggqvist's Counterfactual Refuter
Systems Reply	Reject 3.	Reject iii.
Robot Reply	Reject 1.	Reject i.
Brain Simulator Reply	Reject 1.	Reject i.
Combination Reply	Reject both 1. and 3.	Reject both i. and iii.
Many Mansions Reply	Reject 1.	Reject i.
Bad Intuition Reply	Reject 4.	Reject iv.
Non-human Ability Reply	Reject 5.	Reject ii.
Anti-computationalism Reply	Reject 1.	Reject i.

I. Sorensen & Häggqvist

II. Searle's Chinese Room

III. (In)consistency

IV. Moving forward

The schemas are *not* inconsistent

Sorensen's (and Häggqvist's) schemas are *not inconsistent!*

The Necessity Refuter is *consistent* for the exceptional scenario where counterfactual $\Box \rightarrow$ is **vacuously true**

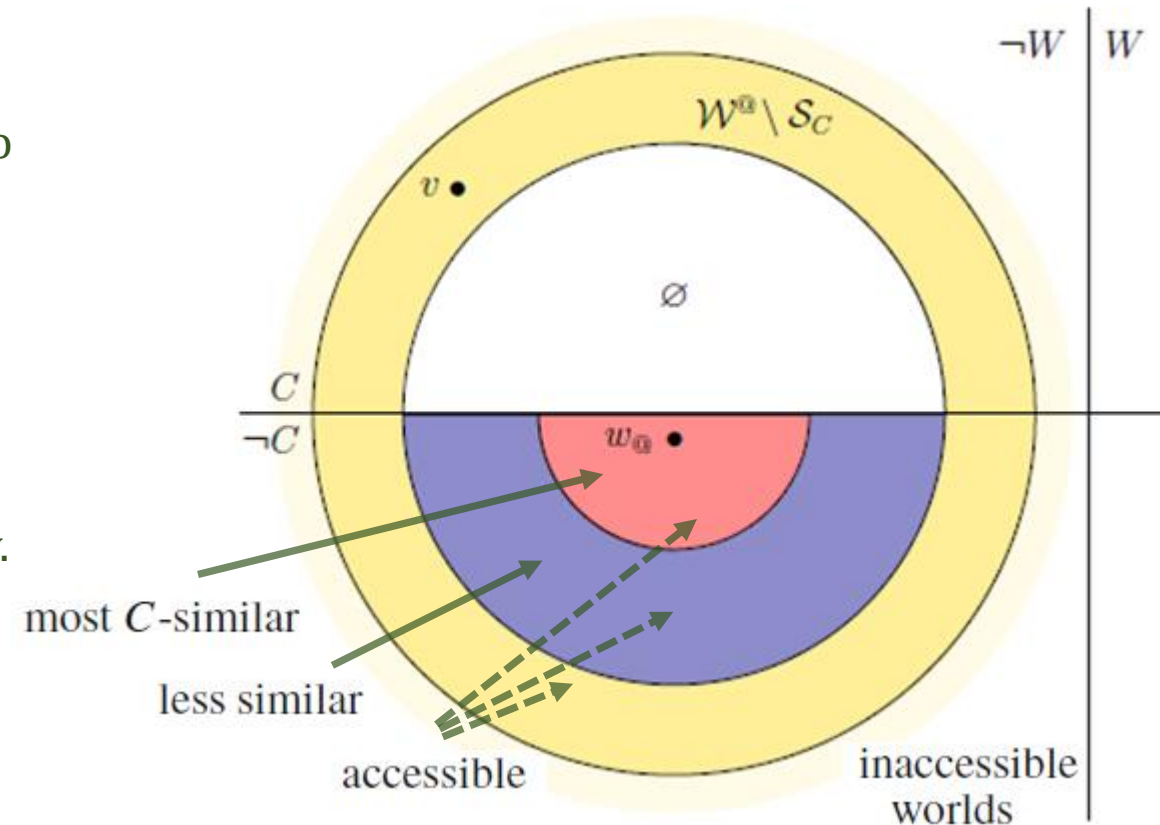
Let's attempt an inconsistency proof:

- from 1. and 2. we obtain $\Box I$;
- from $\Box I$ and 5. we obtain $\Diamond(I \& C)$;
- from here: $\Diamond(I \& C)$ and 3. to conclude $\Diamond W$,
because this would contradict 4., resulting in inconsistency.

Right?

→ No! (c) assumes ' $\Box \rightarrow$ ' is substantively true!

But it can also be vacuously true: no similar C-worlds



Making the schemas properly inconsistent

To fix this, add:

$$6. (\Box I \ \& \ \Diamond C) \rightarrow \Diamond(I \ \& \ C)$$

$$7. (\Diamond(I \ \& \ C) \ \& \ (I \ \& \ C) \Box \rightarrow W) \rightarrow \Diamond W$$

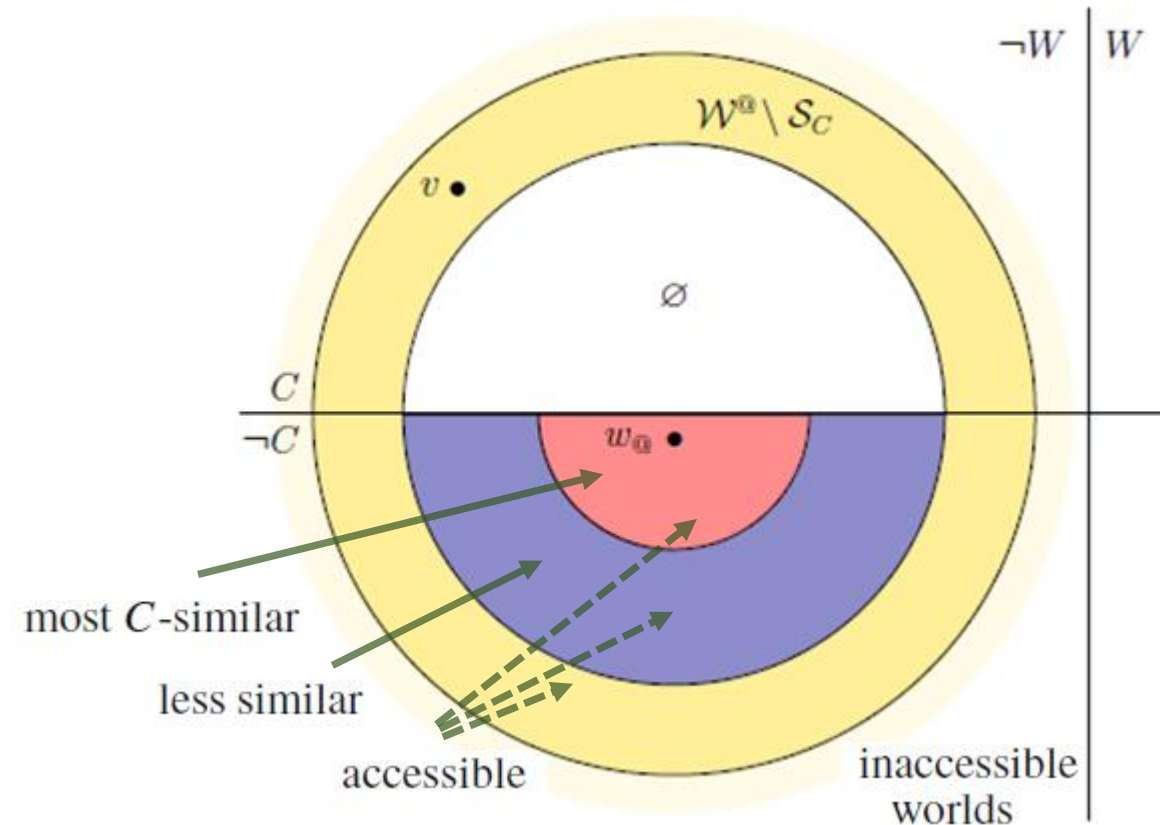
Premise 6. turns out to be a logical truth

Premise 7. is a genuine addition

In principle: a new route to respond to thought experiments,
But perhaps this is merely logic chopping?

Philosophical interpretation: accessible worlds are not sufficiently *similar* even though they are *accessible*

Close to opposing intuitions about ‘counterpossibles’
(Williamson 2017, Berto *et al.* 2017, Sendłak 2019)



I. Sorensen & Häggqvist

II. Searle's Chinese Room

III. (In)consistency

IV. Moving forward

Further applications of the schemas

There exist so many controversies over destructive thought experiments!

If you know the debate well: *lots of low-hanging fruit!*

Successful application of the schemas creates conceptual clarity:
why do disagreeers disagree?

I know of just two worked-out papers applying the schemas:

1. Damper (2006): Searle's CRA (Sorensen's schemas)
2. Linsbichler & Da Cunha (March 2023):
Otto Neurath's Utopia (Häggqvist's schema)
3. ... you?
(e.g. Gould's "Replaying Life's Tape"; or the Hayden–Preskill protocol for
black hole radiation, ...)



Bibliography

Francesco Berto, Rohan French, Graham Priest, David Ripley (2017). Williamson on Counterpossibles. *Journal of Philosophical Logic* **47**, pp. 693–713. [[available online](#)]

Elke Brendel (2017). The Argument View: Are thought experiments mere picturesque arguments? *The Routledge Companion to Thought Experiments*. Imprint Routledge. Edited Michael T Stuart, Yiftach Fehige, James Robert Brown. [[available online](#)]

Robert Damper (2006). The logic of Searle's Chinese room argument. *Mind and Machines* **16**, pp. 163–183. [[available online](#)]

Sören Häggqvist (2009). A Model for Thought Experiments. *Canadian Journal of Philosophy* **39** (1), pp. 55–76. [[available online](#)]

David Lewis (1973). *Counterfactuals*. Malden, Massachusetts: Blackwell Publishers Inc.

Alexander Linsbichler & Ferreira Da Cunha (March 2023). Otto Neurath's Scientific Utopianism Revisited-A Refined Model for Utopias in Thought Experiments. *Journal for General Philosophy of Science* **54**, pp. 233–258. [[available online](#)]

Ruward Mulder & F.A. Muller (Jan. 2023). Modal-Logical Reconstructions of Thought Experiments. *Erkenntnis* **2023**. [[available online](#)]

John Norton (2004). On thought experiments: Is there more to the argument? *Philosophy of Science* **71**, pp. 1139–1151. [[available online](#)]

John Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences* **3**, pp. 417–57 [[available online](#)]

Maziej Sendłak (2019). On the Pragmatic Approach to Counterpossibles. *Philosophia* **47**, pp. 523–532. [[available online](#)]

Roy A. Sorensen (1992). *Thought experiments*. Oxford University Press.

Michael Stuart (2018). How Thought Experiments Increase Understanding. *The Routledge Companion to Thought Experiments*, pp. 526-544. Imprint Routledge. Edited Michael T Stuart, Yiftach Fehige, James Robert Brown. [[available online](#)]

Mauricio Suárez (forthcoming). The Representations in Thought Experiments.

Timothy Williamson (2017). Counterpossibles in Semantics and Metaphysics. *Argumenta* **2** (2), pp. 195–226. [[available online](#)]

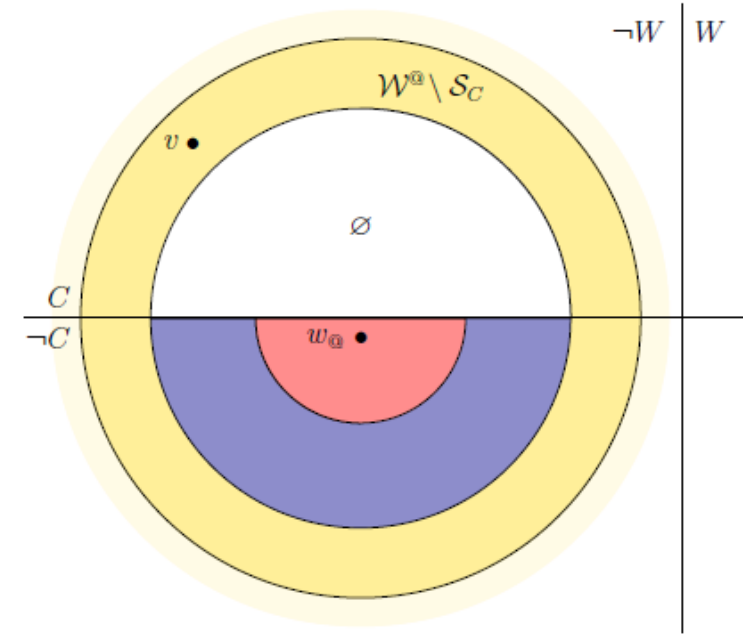
Häggqvist's modal schema applied to CRA

- | | |
|---|---|
| i. T | i. The theory of strong A.I. a computer passing the Turing test understands. |
| ii. $\diamond C$ | ii. It is possible that Searle hand-implements the algorithm. |
| iii. $T \rightarrow (C \Box \rightarrow W)$ | iii. Strong A.I. implies that were Searle to hand-implement the algorithm, he would understand. |
| iv. $(C \Box \rightarrow \neg W)$ | iv. Were Searle to hand-implement the algorithm, he would not understand. |

Truth conditions for the Necessity Refuter: consistent!

Exceptional situation where both $\mathcal{S}_C \cap \mathcal{W}_C^@ = \emptyset$ and $\mathcal{W}_W^@ = \emptyset$; hence $C \Box \rightarrow W$ is true because $\emptyset \subseteq \emptyset$. W is false in every accessible world, but among the dissimilar worlds there are W -worlds: to the right of the vertical line subdividing all worlds $\mathcal{W} \supset \mathcal{W}^@$ into \mathcal{W}_W and $\mathcal{W}_{\neg W}$. This situation depicted makes all premises of Sorensen's Necessity Refuter true, thereby showing their *consistency*

- (tc1) $w_@ \models S$ iff $w_@ \in \mathcal{W}_S^@$.
- (tc2) $w_@ \models S \rightarrow \Box I$ iff $w_@ \notin \mathcal{W}_S^@$ or $\mathcal{W}_I^@ = \mathcal{W}^@$.
- (tc3) $w_@ \models (I \wedge C) \Box \rightarrow W$ iff $\mathcal{S}_{I \wedge C} \cap \mathcal{W}_{I \wedge C}^@ = \emptyset$ or, for some k :
 $(\mathcal{S}_{I \wedge C}^k \cap \mathcal{W}_{I \wedge C}^@) \cap \mathcal{W}_{\neg W}^@ = \emptyset$ and $\emptyset \subset \mathcal{S}_{I \wedge C}^k$
- (tc4) $w_@ \models \neg \Diamond W$ iff $\mathcal{W}^@ = \mathcal{W}_{\neg W}^@$.
- (tc5) $w_@ \models \Diamond C$ iff $\emptyset \subset \mathcal{W}_C^@$.



Inconsistency of the Necessity Refuter!

Sorensen's Necessity Refuter:

1. T
2. $T \rightarrow \Box I$
3. $(I \& C) \Box \rightarrow W$
4. $\neg \Diamond W$
5. $\Diamond C$

Attempted Proof of Inconsistency. From 1 and 2 we obtain $\Box I$ by *modus ponens*. Combine $\Box I$ with 5 ($\Diamond C$) to obtain $\Diamond(I \wedge C)$. To prove an inconsistency, one would like to proceed from here that the antecedent of 3 ($I \wedge C \Box \rightarrow W$) is satisfied, such that it follows that $\Diamond W$, which would contradict 4 ($\neg \Diamond C$). However, in this last step, one has surreptitiously assumed the counterfactual $\Box \rightarrow$ to be substantively true, forgetting it can also be vacuously true. The attempted proof fails. We shall now proceed to show the consistency explicitly.

Inconsistency of the Necessity Refuter!

Sorensen's adjusted Necessity Refuter:

1. T
2. $T \rightarrow \Box I$
3. $(I \& C) \Box \rightarrow W$
4. $\neg \Diamond W$
5. $\Diamond C$
6. $(\Box I \& \Diamond C) \rightarrow \Diamond(I \& C)$ (logical truth)
7. $(\Diamond(I \& C) \& (I \& C) \Box \rightarrow W) \rightarrow \Diamond W$

Proof the Inconsistency of 1–7. From premises 1 and 2 we obtain $\Box I$ by *modus ponens*. Combine $\Box I$ with premise 5 ($\Diamond C$) to obtain the antecedent of premise 6, and then, again by *modus ponens*, we obtain $\Diamond(I \wedge C)$. From the conjunction with premise 3, we obtain the antecedent of premise 7, and then, again by *modus ponens*, we deduce $\Diamond W$, which contradicts premise 4 ($\neg \Diamond W$). *Q.e.d.*

Standard Reply to the Chinese Room Argument (I/II)

- (a) *Systems Reply*: the whole room understands Chinese, not just Searle alone, who merely functions as a part of it.
- (b) *Robot Reply*: if one were to provide the program α with causal powers in the form of a body to operate, i.e. an embodied program, like a robot, equipped with an electric system of sensors and effectors, then it would understand Chinese.
- (c) *Brain Simulator Reply*: a particular computer program, mimicking one-to-one the behaviour of neurons and synapses of a human Chinese-understanding brain, will understand Chinese.
- (d) *Combination Reply*: A combination of the systems reply, brain simulator reply and robot reply add up to a Chinese-understanding computer program, through the help of its body, part of which exactly mimics the structure of a Chinese-understanding brain.

Standard Reply to the Chinese Room Argument (II/II)

- (e) *Other Minds Reply*: Because there exist computer programs that put out the same behaviour as human beings, and because behaviour is the sole reason that we contribute understanding to other human beings, we must contribute understanding to such programs.
- (f) *Many Mansions Reply*: Since there are so many different kinds of computers imaginable, even though current technology is lacking, we are bound up to build precisely that kind of computer that simulates the right kind of conditions to produce mental states such as understanding.
- (g) *“Bad Intuition Reply”*: Despite Searle’s insistence that he does not understand Chinese, this is only his intuition. In a situation like the CRA, it is likely that he *will* understand Chinese.
- (h) *“Non-human Ability Reply”*: The hand-implementation of α is – even in principle – too complicated for a human being to do. If Searle were nevertheless able to do this, he will cease to be human.
- (i) *“Anti-computationalism Reply”*: By focusing on symbol manipulation only, Searle attacks a strawman version of AI, namely computationalism. True proponents of AI would hold that the program involves more than that.