What is Lewisean Interpretivism?

Edward Elliott

School of Philosophy, Religion and History of Science
University of Leeds

e.j.r.elliott@leeds.ac.uk

18 June 2019

Two Principles of Interpretation

'Radical Interpretation': Lewis' account of *interpretation*, based on two key principles:

Principle of Rationalisation

Assign beliefs and basic desires to an agent so as to maximise their *pragmatic rationality*, given the facts about their behaviour

Principle of Charity

Assign an agent (i) beliefs so as to maximise their *epistemic rationality* given the facts about the agent's life history of evidence; and (ii) basic desires appropriate for the kind of being the agent is

Exegetical Questions

Question 1: *How do these two principles interact?* What are we to do when they pull in different directions?

Example: Monsters in the closet

Karl has no significant evidence that there's a monster in his closet, and overwhelming evidence that there isn't; however, he acts in a way that would clearly be well explained by high confidence in the existence of closet-dwelling monsters (and a desire to not be eaten).

Exegetical Questions

The interaction question crops up in, *inter alia*,

- (Maher 1993, pp. 31-2)
- (Hájek & Eriksson 2007, pp. 200-1)
- (Hattiangadi 2019)

Obvious aggregation strategies...

- *Tradeoffs* — which 'weightings' to use, and why?
- *Lexical priority* — which gets priority, and why?

Exegetical Questions

Quote: 'Radical Interpretation', p. 341

First step: using [the physical facts] both as a source of information on Karl's behaviour and as a source of information on his life history of evidence, fill in [his beliefs and desires] completely by means of the Rationalisation Principle and the Principle of Charity.

Question 2: *Why didn't Lewis say more about this?*

Overview

In the remainder, I will argue...

- Question 1:
  Lewis held a *lexical priority* view, but between *causal* and *non-causal* principles of interpretation
- Question 2:
  Lewis didn't discuss the interaction because he (rightly) considered the issue to be of minor importance

Postscripts to 'Radical Interpretation'

A Quick Detour...

The discussion of interpretation in 'Radical Interpretation' is misleading in two important (and related) respects:

- It is *individualistic*
- It describes the interpretation of *persons*, not *states*

Interpretation of States and Individuals

Quote: 'Postscripts to Radical Interpretation', p. 119

I stated my problem in an unduly individualistic way: give the facts *about Karl* as a physical system, solve for the facts *about him* as a person—*his* beliefs, desires, and meanings...

In 'Mad Pain and Martian Pain', I argued that a 'madman' might be in pain not because his state occupied the causal role of pain in him but rather because that state occupies that role, for the most part, in members of the kind to which he belongs. The same possibility should be recognised for attitudes as well.

Interpretation of States and Individuals

The important points:

1. *Interpretation* involves first assigning contents to *states*
   - These are "most likely states of the brain" (1983, p. 373)
   - In later works, whole systems of (degrees of) belief and basic desires $\langle \mathcal{B}el, \mathcal{D}es \rangle$ are assigned to *total physical states*
2. We're to interpret *states* in such a way as to maximise Charity and Rationalisation for the population *overall*
   - Later treated as equivalent to: the interpretation of a state depends on its *typical causal role*
3. Interpretation of *individuals* depends on what states they're in

Reinterpreting 'Radical Interpretation'

Rationalisation in 'Radical Interpretation'

Quote: 'Radical Interpretation', pp. 337–8

Karl should be represented as a rational agent; the beliefs and desires ascribed to him should be such as to provide good reasons for his behaviour, as given in physical terms...

I would hope to spell this out in decision-theoretic terms, as follows. Take a suitable set of mutually exclusive and jointly exhaustive propositions about Karl's behaviour at any given time; of these alternatives, the one that comes true according to $\mathcal{P}$ should be the one (or: one of the ones) with maximum expected utility according to the total system of beliefs and desires [$\langle \mathcal{B}el, \mathcal{D}es \rangle$] ascribed to Karl at that time.

Rationalisation in 'Radical Interpretation'

Causal Rationalisation

An interpretation $\langle \mathcal{B}el, \mathcal{D}es \rangle$ **$\mathcal{R}$-fits** a total physical state S iff...

the typical agent in S will behave in way B that maximises expected utility w.r.t $\langle \mathcal{B}el, \mathcal{D}es \rangle$

'$\mathcal{R}$-*fit*' corresponds to a forwards-looking typical causal role:

$$\langle \mathcal{B}el, \mathcal{D}es \rangle \Rightarrow B, \quad \text{where } B \text{ maximises EU w.r.t. } \langle \mathcal{B}el, \mathcal{D}es \rangle$$

Charity in 'Radical Interpretation'

Principle of Charity

Assign an agent (i) beliefs so as to maximise their *epistemic rationality* given the facts about the agent's life history of evidence; and (ii) basic desires appropriate for the kind of being the agent is

In a Bayesian framework, part (i) requires:

- $\mathcal{B}el$ is probabilistically coherent
- $\mathcal{B}el$ updates by conditionalisation

It'll be helpful to break 'Charity' down into causal/non-causal parts...

Charity in 'Radical Interpretation'

Let $\mathcal{B}el_E = \mathcal{B}el$ conditionalised on E

Causal Charity

Interpretations $\langle \mathcal{B}el, \mathcal{D}es \rangle$ and $\langle \mathcal{B}el_E, \mathcal{D}es \rangle$ **$\mathcal{C}$-fit** states S and $S^\star$ iff...

the typical agent in state S when presented with a stream of evidence E will come to be in state $S^\star$

- '$\mathcal{C}$-*fit*' corresponds to a forwards-/backwards-looking typical role:

$$\langle \mathcal{B}el, \mathcal{D}es \rangle + E \Rightarrow \langle \mathcal{B}el_E, \mathcal{D}es \rangle$$

Charity in 'Radical Interpretation'

Non-Causal Charity

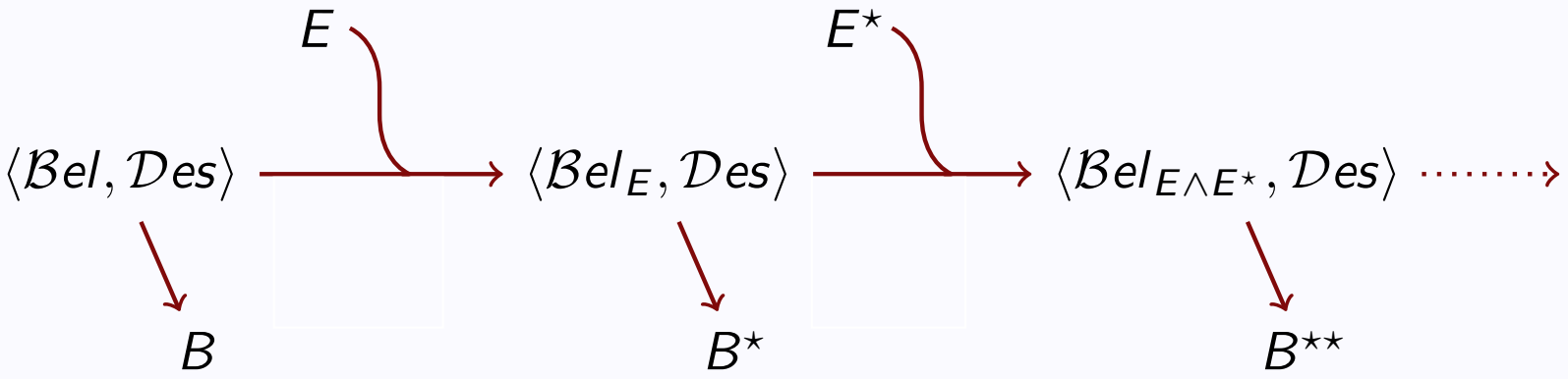
An interpretation $\langle \mathcal{B}el, \mathcal{D}es \rangle$ is *better* to the extent that $\mathcal{B}el$ and $\mathcal{D}es$ are *reasonable* relative to the agent's kind

- Non-causal — depends on...
  1. intrinsic characteristics of the interpretation
  2. facts about the agent's kind

Combining Charity and Rationalisation

- **Causal Rationalisation**
  $\langle \mathcal{B}el, \mathcal{D}es \rangle \Rightarrow B$, where B maximises EU w.r.t. $\langle \mathcal{B}el, \mathcal{D}es \rangle$
- **Causal Charity**
  $\langle \mathcal{B}el, \mathcal{D}es \rangle + E \Rightarrow \langle \mathcal{B}el_E, \mathcal{D}es \rangle$

For any stream of evidence $E, E^\star, \ldots$, and any starting point $\langle \mathcal{B}el, \mathcal{D}es \rangle$:



'New Work for a Theory of Universals'

The Role of 'Charity'

Quote: 'New Work for a Theory of Universals', p. 375

If we rely on principles of fit to do the whole job, we can expect radical indeterminacy of interpretation. We need further constraints, of the sort called principles of (sophisticated) charity, or of 'humanity'. Such principles call for interpretations according to which the subject has attitudes that we would deem reasonable for one who has lived the life that he has lived...

These principles select among conflicting interpretations that equally well conform to the principles of fit.

Principles of Fit

Quote: 'New Work for a Theory of Universals', p. 374

Say that $\mathcal{B}el$ and $\mathcal{D}es$ *rationalise* behaviour B after evidence E iff the system of desires given by the $\mathcal{B}el_E$-expectations of $\mathcal{D}es$ ranks B at least as high as any alternative behaviour.

Say that $\mathcal{B}el$ and $\mathcal{D}es$ *fit* iff, for any evidence-specifying E, [learning] E yields a state that would cause behaviour rationalised by $\mathcal{B}el$ and $\mathcal{D}es$ after E.

That is our only constraining principle of fit.

Principles of Fit

Principle of Fit

An interpretation $\langle \mathcal{B}el, \mathcal{D}es \rangle$ **$\mathcal{F}$-fits** a total physical state S iff...

for any evidence E, the typical agent in S when presented with a stream of evidence E will come to be in a state $S^\star$ that causes some behaviour B which maximises expected utility with respect to $\langle \mathcal{B}el_E, \mathcal{D}es \rangle$

- '*$\mathcal{F}$-fit*' is supposed to help capture both:
  - the role of evidence in the adjustment of beliefs
  - the role of beliefs/desires in the production of behaviour
- However, it is entirely *forwards-looking*:

$$\langle \mathcal{B}el, \mathcal{D}es \rangle + E \Rightarrow S^\star \Rightarrow B, \quad \text{where } B \text{ maximises EU w.r.t. } \langle \mathcal{B}el_E, \mathcal{D}es \rangle$$

Principles of Fit

Quote: 'New Work for a Theory of Universals', p. 374

That is our only constraining principle of fit. (Where did the others go? — *We built them into the definitions whereby $\mathcal{B}el$ and $\mathcal{D}es$ encapsulate an assignment of content to various states.*)

Principles of Fit

Quote: 'New Work for a Theory of Universals', p. 374

$\mathcal{B}el$ is a probability distribution... regarded as encapsulating the subject's dispositions to form beliefs under... evidence: if a stream of evidence specified by proposition E would put the subject into a total state S—for short, *E yields S—we interpret S to consist in part of the belief system... that comes from $\mathcal{B}el$ by conditionalisation on E*.

$\mathcal{D}es$ is... regarded as encapsulating the subject's basic desires: if E yields S, *we interpret S to consist in part of the system of desires given by the $\mathcal{B}el_E$-expectations of $\mathcal{D}es$*.

Principles of Fit

Forwards Coherence

S can be interpreted $\langle \mathcal{B}el, \mathcal{D}es \rangle$ only if...

for any stream of evidence E, if $S + E \Rightarrow S^\star$, then $S^\star$ is interpreted $\langle \mathcal{B}el_E, \mathcal{D}es \rangle$

Backwards Coherence

S can be interpreted $\langle \mathcal{B}el, \mathcal{D}es \rangle$ only if...

for any stream of evidence E, if there exists an $S^\star$ such that $S^\star + E \Rightarrow S$, then $S^\star$ is interpreted in some way $\langle \mathcal{B}el^\star, \mathcal{D}es \rangle$ such that $\mathcal{B}el^\star_E = \mathcal{B}el$

Principles of Fit

Fit + Backwards Coherence + Forwards Coherence
=
General Principle of Fit

General Principle of Fit

An interpretation $\langle \mathcal{B}el, \mathcal{D}es \rangle$ **fits** a total physical state S iff...

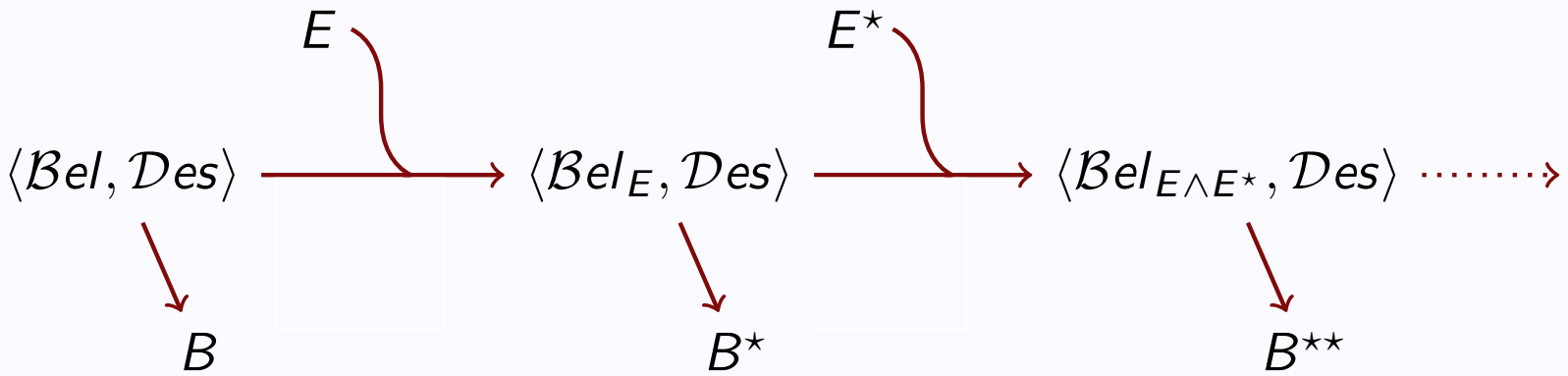
for any stream of evidence E,

- $\langle \mathcal{B}el, \mathcal{D}es \rangle$ $\mathcal{F}$-fits S
- if $S + E \Rightarrow S^\star$, then $\langle \mathcal{B}el_E, \mathcal{D}es \rangle$ $\mathcal{F}$-fits $S^\star$
- if there exists an $S^\star$ such that $S^\star + E \Rightarrow S$, then some $\langle \mathcal{B}el^\star, \mathcal{D}es \rangle$ such that $\mathcal{B}el^\star_E = \mathcal{B}el$ $\mathcal{F}$-fits $S^\star$

The Structure of Lewisean Interpretivism

General Principle of Fit
=
Causal Rationalisation + Causal Charity

For any stream of evidence $E, E^\star, \ldots$, and any starting point $\langle \mathcal{B}el, \mathcal{D}es \rangle$:



Answering the Exegetical Questions

Question 1:

Fit with functional role takes lexical priority over considerations of 'reasonableness', but both Rationalisation and (a part of) Charity are involved in characterising that role

Question 2:

The interpretation of states depends on what they do in the *typical* case — and plausibly, as a matter of fact in the typical case Rationalisation and Charity don't pull in different directions

Agents who don't behave in a manner that seems pragmatically rational given what would be epistemically rational in light of their evidence are *deviant cases* — the normal rules don't apply

Thank you

Thanks to Jessica Isserow for comments on a practice run!

Quote: *On the Plurality of Worlds*, p. 107

Such a theory, I said, should have two parts. *One part says what it is for an assignment of content to states to fit the functional roles of the states*; the constraints are principles of rationality, for instance a principle to the effect that a state which is assigned content consisting of some system of beliefs and desires ought to be a state that tends to produce conduct that would serve those desires according to those beliefs. But principles of fit can be expected to underdetermine the assignment of content very badly... *Therefore a theory of content needs a second part: as well as principles of fit, we need 'principles of humanity'*, which create a presumption in favour of some sorts of content and against others.