FICTIONS, MODELS, THOUGHT EXPERIMENTS

Julian Reiss

What if?
THOUGHT EXPERIMENTS IN HISTORY
Julian Reiss
This paper will consider what-if fictions (‘whiffs’) in history as test case for epistemological claims that have been made about thought experiments (‘TEs’)

To proceed I will argue, first, that whiffs are TEs and, second, give some reasons for believing that they are special kinds of TEs with properties that make them particularly interesting for epistemic purposes

I will then apply major epistemological accounts of TEs and show that they are insufficient to show how we learn from whiffs

In response, I outline an alternative epistemology that applies successfully to whiffs

I conclude by saying that the account defended here preserves what’s good about its major alternatives
SOME EXAMPLES

- The genre of historical ‘what-if’ fiction has a long pedigree and experienced a recent explosion:
  - *If It Happened Otherwise: Lapses into Imaginary History* (1932)
  - *If I Had Been: Ten Historical Fantasies* (1979)
  - *Virtual History: Alternatives and Counterfactuals* (1997)

- Many of these have mostly recreational value (most prominently, perhaps Robert Harris’s *Fatherland*) but others have cognitive functions:
SOME EXAMPLES

- Usually the starting point of such ‘whiffs’ is a counterfactual in which an event of historical interest (often: decisions taken by political leaders the continued life or death of central figures) is removed from history.

- Then a story is told how events would have unfolded under the counterfactual antecedent but using connecting principles that are (believed to be) true of the actual world.

- Thus:
  - If, counterfactually, the Athenians had lost the battle at Salamis, the rise of Western values would have been very unlikely (behavioural generalisation).
  - If, counterfactually, the UK had confronted Hitler over the Sudentenland, WWII would have been no worse as either Hitler would have backed down, a coup d’état would have been staged or the war would have broken out earlier (particular evidence).
  - If, counterfactually, the railroad would not have been introduced in the U.S. in the 19th century, economic growth would have been only minimally different (theory).
Much ink has been spilled on developing semantics for such counterfactuals that focus for instance on

- what is an acceptable counterfactual antecedent and how is it to be implemented?
- how is the counterfactual consequent to follow from the antecedent?

See for instance Tetlock and Belkin 1996; Tetlock, Lebow and Parker 2006

This topic I’ve discussed elsewhere (PSA 2008)

Here I want to draw some lessons from this debate for TEs in general

To do so I first have to establish that whiffs are TEs
ARE WHIFFS THOUGHT EXPERIMENTS?

- Rather than defining a TE in general and then showing that whiffs fall under the definition, what I want to do here is show that the two practises share a great number of salient characteristics, a fact which indicates that they are related.

- Trying to find an appropriate definition could easily be question begging.

- Moreover, many concepts are best seen as cluster concept, and TEs are no different.
HYPOTHETICAL NATURE

- e.g., Norton 1991: ‘Thought experiments are arguments which: (i) posit hypothetical or counterfactual states of affairs...’

- The counterfactual nature of whiffs is evident and explicit

- But: unlike some TEs, they do not posit impossible states of affairs (to the contrary: much effort is spent on demonstrating that the antecedent is ‘historically consistent’ or realistic – false but possible)

- It is thus preferable to say ‘hypothetical’ in both cases

- Brown 2004: ‘The so-called counterfactual nature of [TEs] is overstressed. And... so is idealization’

- In both cases there wouldn’t be much point in performing the exercise
EXECUTED IN THE MIND

- e.g., Sorensen 1992: ‘A TE is an experiment... that purports to achieve its aim without the benefit of execution’

- Trivially, the same is true of whiffs: it is neither possible nor desirable to implement/instantiate a whiff

- This is probably the only necessary condition – an RCT isn’t a thought experiment
Several philosophers have emphasised the important role the ‘mind’s eye’ plays in TEs

e.g., Brown 2004, Sorensen 1992, de Mey 2006

The importance of historical imagination is widely recognised and discussed (most prominently perhaps in Collingwood; see also Trevor-Roper: History and Imagination and Scriven 1966), so I won’t pursue the topic here, except to note: ‘One [the skilled historian] “sees” the explanation via verstehen – but the act of “seeing” is a highly tested skill, as is “seeing” the solution of a bridge or chess problem, or “seeing” that a set of tracks are those of a red fox running’ (Scriven)
Almost the entire community of philosophers of the TE accepts that TEs don’t produce new empirical data (an important exception: McAllister).

And of course the same is true of whiffs: they are constructed on the basis of evidence but don’t produce new – comparable – evidence (such as historical records).
FUNCTION: ALETHIC REFUTERS

- And now for some more interesting claims about TEs and whiffs
- Like TEs (e.g., Sorensen 1992), whiffs often function as necessity refuters by showing that a certain ‘inevitable’ outcome could in fact have been prevented if only another event had taken place (and the whiff shows that that was possible)
- The necessity statements often come from an accepted theory:
  - nuclear deterrence theory
  - neo-realist balancing theory
  - ‘triumphalism’
- E.g., $S$ (nuclear deterrence theory); $S \supset □I$ (peaceful resolution of the Cuba crisis); $(I \& C) □ → W$ – ‘If peace were to prevail and the U.S. would have launched air strikes against the Soviet bases then the USSR would not have retaliated’; $\sim ◇ W$ (‘that seems absurd’); ◇C (a whiff shows that C was a real possibility: if the Soviets hadn’t withdrawn, the U.S. would have struck)
Cate Elgin has argued that TEs sometimes function by reconfiguring previously held beliefs (2002; 2007).

The idea is two-fold: (a) TEs teach us about possibilities that we could have contemplated but for some reason or another didn’t; (b) learning about such possibilities constitutes genuine cognitive advancement.

This is precisely an important role of whiffs: the need for explanatory closure and covering laws and the hindsight bias are an important cognitive biases among historians, and whiffs help them to see that certain events were ‘objectively possible’ (though without providing new evidence), see Tetlock and Lebow 2001.
FUNCTION: NEW CAUSAL KNOWLEDGE

- TEs often teach us something genuinely new about causal relations (such as the causal independence of rate of fall and weight)

- And this, of course, is the most important function of what; they are used to (Bulhof 1999):
  - explain singular historical outcomes
  - attribute responsibility
  - determine the ‘historical significance’ of certain events

- These are all related to singular causation; but they also teach us something general: that a certain type of factor has the causal capacity to affect an outcome in a specific way
Several authors have emphasised that TEs and fictions (of a certain kind) are closely related:

- e.g., by arguing that fictions are TEs (Elgin 2007; Carroll 2007)
- or, conversely, by arguing that TEs are fictions (Davies 2007; cf. Sorensen 1992)

There is certainly continuity between a ‘scientific’ whiff and a literary historical fiction.
These and more features are therefore shared between TEs and whiffs, which is enough reason to believe that they are closely related.

But whiffs have some additional characteristics, which in my view make them particularly interesting as test cases for general theories of thought experimentation:

- Like a scientific TE, whiffs are fully responsive to empirical evidence and changes in theory.
- And like a philosophical TE, it would be utterly futile to try to execute a whiff.

Thus: a whiff result is far from arbitrary; but it cannot be accessed via independent means.

(Background: Peijnenburg and Atkinson have recently argued that philosophical TEs are often 'bad ones' in that they beg the question or have contradictory results; scientific TEs can be bad but disagreements can be resolved because of new theories or new experiments.)
I take the major accounts of TE to be well-known here, so let me merely point out why I think three dominant accounts fail:

- Brown’s Platonism
- Norton’s Empiricism
- Gendler’s constructivism
PLATONISM

- Prima facie, whiffs are very good candidates for Platonism:
  - ‘First, although it is true that empirical knowledge is present in this example, there are no new empirical data being used when we move from Aristotle’s to Galileo’s theory of free fall. And second, it is not a logical truth’ (Brown 2004)
  - Moreover, they frequently are ‘Platonic TEs’ in that they simultaneously refute an old theory (explanation) and establish a new theory (explanation)
  - But a major ingredient in Brown’s Platonism is the intuition of an abstract relation of nomic necessity, and whatever the mind’s eye ‘sees’ when contemplating a whiff, it is not the causal relation at stake
  - Instead, that causal relation is inferred on the basis of the whiff and knowledge about the actual course of history (cf. a ‘natural experiment’)
Norton claims that ‘If [TEs] can be used reliably epistemically, then they must be arguments (construed very broadly) that justify their outcomes or are reconstructible as such arguments’ (2004)

I just said that establishing a causal claim on the basis of a whiff involves an inference – isn’t that just inductive logic?

There are two ways to apply Norton’s principle to history:

- Demand that all historical explanations are D-N (as did for instance E.H. Carr); this is now widely rejected
- Claim that the counterfactual consequent deductively follows from the antecedent and suitable facts and laws (and then apply inductive logic to the final step)
  - This is just Goodman’s account of counterfactuals and that has problems in its own right
  - Principles that are strong enough to entail typical consequents are few and far between in history
    - Methodological principles used: ‘consistency with known laws, facts etc.’
- Moreover, don’t we confirm general causal claims by instances?
CONSTRUCTIVISM

- By contrast, I think that constructivism and its idea of reconfiguration faithfully describes one function of TEs

- Idea: ‘... the Aristotelian comes to have novel justified true beliefs about the empirical world... because he has performed an act of introspection that brings to light heretofore unarticulated and (because he lacked a theoretical framework in which to make sense of them [or, in our case more specifically: because he had adopted a theoretical framework that wouldn’t allow him to make sense of them]) heretofore implausible tacit beliefs’; and ‘[the Aristotelian is led] to a reconfiguration of his conceptual commitments of a kind that lets him see familiar phenomena in a novel way’ (Gendler 1998)

- However, it leaves out the more important role, viz., the establishment of new singular causal claims

- Here I’m with Norton 2004: ‘... whatever its merits may be, [constructivism] cannot supply a complete epistemology of thought experiments in science’ – nor of TEs in history
AN ANALOGY

- When trying to establish the reliability of a TE or whiff, we are in the same epistemic situation as the experimenter who doesn’t know whether a measurement instrument is reliable but cannot access the property of interest directly.

- This can be formulated as a regress: to know whether we have a reliable measurement instrument, we need to know whether it produces the correct outcome; but to know whether it produces the correct outcome we need to know whether it is reliable.

- Luckily, there are well-known experimentalist strategies that help to improve reliability.
The history of science is replete with examples of types of error that can be made when establishing an experimental result or using the result to infer a claim about a target system of interest.

Claim: an experimental procedure is reliable to the extent that errors in establishing the experimental result or inferring claims about target systems of interest have been eliminated, controlled or corrected for.

At this level of generality little else can be said – knowledge about potential sources of error are highly context specific.

Note: if Norton’s notion ‘logic’ is broad enough as include this error-correction model as a kind of inductive logic, this approach collapses into his (but is this really logic?)
Clearly, it is historical expertise that drives the result, not ‘novice intuition’; but expertise isn’t enough.

Here are some examples of known errors that experts are routinely subject to:

- Hindsight bias
- Framing effects
- Conjunction fallacy
- Overemphasis of deviations from the normal
- Theory bias
- Various motivational biases

Another strategy is robustness: use types of TE that tend to produce robust results (one way to justify the minimal-rewrite rule)
One significant difference between a whiff and a scientific TE is the major epistemic challenge for the former lies in establishing its internal validity (IV; and there isn’t a problem of external validity: EV), for the latter it lies in establishing EV (and, rightly or wrongly, the problem of IV is put aside).

This difference is easily accommodated within my account: the experimental strategies to establish EV are no different in principle from those that establish IV.

That some physics TEs are regarded as successes depends on two characteristics of causes operating in this domain:

- Additivity
- Swamping

In domains where causal effects are highly context sensitive and small, TEs are unlikely to play an important role:

- Without additivity it does not make sense to ask what factors do ‘all by themselves’
- Without additivity and swamping, the computational demands would be likely to exceed the brain’s computational power

In these domains we have related methods such as computer simulation that can substitute TEs.
There is a life debate in the philosophy of history over the ineliminability of counterfactuals and thus TEs in history

- Many historians are still influenced by positivism and hence abhor them
- Others claim that we can’t get rid of them save at the expense of exceedingly dull descriptions of unrelated sequences of events

- While I don’t have a problem with TEs, I don’t think that the use of causal language always implicitly refers to counterfactuals
- You can have causation without counterfactuals – and there are associated methods: process tracing
  - Example: European diseases caused settling in the Americas
- Michael Scriven (1966) suggest a method of eliminative induction
CONCLUSION

- The epistemological approach defended here synthesises what’s great about its major alternatives:
  - The central role of the imagination of Brown’s Platonism (plus the somewhat mysterious fact that sometimes we stumble over models/TEs that have external validity)
  - Norton’s empiricist inclinations
  - Gendler’s (and Elgin’s) emphasis of the sometimes constructive role a TE plays