*Perspectives in Science* (forthcoming)

# Fiction as Thought Experiment

**Catherine Z. Elgin**

Jonathan Bennett (1974) maintains that Huckleberry Finn's deliberations about whether to return Jim to slavery affords insight into the tension between sympathy and moral judgment; Miranda Fricker (2007) argues that the trial scene in *To Kill a Mockingbird* affords insight into the nature of testimonial injustice. Neither claims merely that the works prompt an attentive reader to think something new or to change her mind. Rather, they consider the reader cognitively better off for her encounters with the novels. Nor is her cognitive improvement restricted to acquiring new justified true beliefs about the works themselves. What the reader gleans is supposed to enhance her knowledge or understanding of the extra-literary world. Fricker and Bennett are probably right. But their being right raises an epistemological problem. How can a work of fiction which is not and is known not to be true provide any measure of epistemic access to the way things actually are? Is it really possible to find out about the world by making things up?

One attractive answer is that works of fiction are thought experiments. Like literary fictions, thought experiments neither are nor purport to be physically realized. Nevertheless, they evidently enhance understanding of the phenomena they pertain to. If fictions are thought experiments, they advance understanding of the world in the same way that (other) thought experiments do. So we need to ask: 1) How do thought experiments enhance understanding? And 2) are fictions enough like thought experiments that it is reasonable to think that they function in the same way?

In order to answer these questions, it is helpful first to consider how standard experiments

enhance understanding. I suggest that they have more in common with fiction than we ordinarily suppose. I go on to consider thought experiments, and argue that they are in fact small, tightly constrained fictions. Then I look at fiction per se. Throughout, a critical question is how something that does not consist of truths about a range of phenomena can non-accidentally afford epistemic access to those phenomena.

**Experiments**

Reality, as William James said, presents itself as a blooming buzzing confusion (James 1983). Every discernible item has indefinitely many discernible properties and stands in indefinitely many discernible relations. Our only hope of understanding and coping with the confusion that confronts us is to ignore most of what is there to be seen. To a considerable extent, selective disregard is automatic. In order to see anything we overlook a lot. Where automaticity fails, we purposely block things out. We scan the forest by ignoring individual trees, or focus on a tree by screening off the rest of the forest. In some cases, though, more than an act of will is required to selectively disregard what we need to. Then we may resort to experimentation. If it is unfeasible to simply pay no attention to the man behind the curtain, we may contrive a situation from which he is absent.

An experiment is not a mere matter of bringing nature indoors. It is a controlled manipulation of events, designed and executed to make some particular phenomenon salient. Natural entities are multifaceted. Important properties and relations are often masked by the welter of complexities that embed them. So in experimenting, a scientist isolates a phenomenon from many of the forces that typically impinge on it. To the extent possible, she eliminates confounding factors. She holds most ineliminable factors fixed, effectively consigning them to

the cognitive background of things to be taken for granted. This enables effect of the experimental intervention on the remaining variable to stand out. Through such a strategy, she casts into bold relief factors that might typically be hidden from view.

Suppose a population of wild mice who were accidentally exposed to bisphenol-A later exhibited a high rate of liver cancer. To conclude that exposure to bisphenol-A caused their disease would be premature. For all we know, those mice might have been peculiarly susceptible to liver cancer or might have been exposed to a carcinogen we failed to notice. To glean direct, non-anecdotal evidence of a connection between exposure to bisphenol-A and liver cancer, investigators place genetically identical mice in otherwise identical environments, exposing half of them to massive doses of the chemical while leaving the rest unexposed. The common genetic endowment and otherwise identical environments neutralize a multitude of genetic and environmental factors believed to standardly influence the incidence of cancer. This blocks rival explanations that might be proposed for the high rate of cancer in the wild population. If the exposed mice show a significantly higher incidence of cancer than the controls, the difference is apt to be attributed to exposure to bisphenol-A. The attribution is reasonable to the extent that the scientists manufactured a situation where no other explanation of the difference between the target mice and the control group is plausible. Granted, the experiment takes place against a cluster of fallible background assumptions. So it does not afford conclusive evidence. But because of its rigorous controls, it affords stronger and more direct evidence than a mere correlation between exposure and cancer in a wild population would.

Designing an experiment is setting a stage where events can play out. Conducting an experiment involves initiating and perhaps intervening in a course of events. Experiments are dynamic; they unfold over time. Moreover, they unfold as they do at least in part because of a

scientist's actions. She instigates and perhaps interrupts, deflects, impedes, or amplifies a natural sequence. She may isolate phenomena from their normal concomitants and introduce unusual provocations. The mice in the bisphenol-A experiment are exposed to massive doses of the chemical. This allows the scientist to obtain a pronounced effect in a relatively short time. Her working assumption is that a small mammal's exposure to a large dose over a short period is equivalent to a large mammal's exposure to a lower dose over a longer period. If that assumption is true (or close enough), she can safely extrapolate from the laboratory situation to mammals in general, including humans.

Experiments often involve creating and using items that are nowhere to be found in nature. Genetically identical mice are artifacts; their genetic makeup is designed to suit the sorts of experiments in which they will be used. Pure forms of chemicals are artifacts as well, being synthesized under carefully controlled conditions to avoid contamination.

The Miller-Urey experiment begins with inorganic chemicals believed to be present on Earth in prebiotic times. The experiment consists of a sequence of chemical reactions whose ultimate output consists of organic chemicals and amino acids. It thus shows how life could have emerged from non-living matter (Ball 2005). For the experiment to work, the chemicals – methane, ammonia, hydrogen and water – had to be pure. Any hint of contamination would discredit the result. Moreover, to insure that no organic material was accidentally introduced during the course of the experiment, the chemical processes had to be isolated from the environment. Although – indeed because – the experimental components and conditions were unnatural, the experiment revealed something important about the natural world. It did so not by saying that something is the case, but by showing it to be the case. It exemplified a path from inorganic to organic molecules.

To make this out, I need to say a bit about exemplification. Exemplification is the relation of a sample, example, or other exemplar to whatever it is a sample or example of (Goodman 1968; Elgin 1996). A fabric swatch exemplifies an available pattern of cloth; a textbook example exemplifies a mode of reasoning. Exemplification involves a dual referential relationship. An exemplar directly refers to a property or pattern it instantiates or a relation it stands in, and thereby refers indirectly to other items that instantiate that property or pattern, or stand in that relation. An exemplar typifies an extension when it exemplifies a property common to all and only members of that extension. In highlighting, displaying, or making manifest certain of its properties, patterns, and relations, an exemplar affords epistemic access to them. By instantiating and referring to its shade, a splotch on a paint sample card exemplifies teal blue. By typifying teal blue, it aids in the selection of paint. Not being fire engine red, the splotch cannot exemplify fire engine red; it affords no epistemic access to that color.

Exemplification is selective. An exemplar highlights some of its properties by marginalizing or downplaying others. Although the splotch also instantiates a particular shape and size, in its standard use it refers to neither. Whether and what an item exemplifies depends on how it is used. Even though this is not its standard function, the splotch on the sample card could be used in such a way that it exemplified its size, shape, orientation, or distance from Dubuque. By highlighting a property, relation or pattern, the exemplar makes it salient. That enables us to recognize it and appreciate its significance, not only in the exemplar itself, but in the other members of the class it typifies.

In principle any property that is instantiated can be exemplified and any item that instantiates a property can exemplify it. This is why it is so easy to use unassuming, mundane objects as examples. We can just point to a passing car and announce that it is an example of

gas-guzzling arrogance. Assuming that it is an instance of that property, it promptly becomes an example of it. But what is possible is not always practicable. Some properties commingle in such a way that it is hard to point to an instance of one without simultaneously pointing to an instance of the other. To use Quine's example, every creature naturally endowed with a heart is also naturally endowed with kidneys. So every instance of a cordate is an instance of a renate. Arguably, we could get our audience to focus on a creature's having kidneys and ignore the fact it also has a heart, but considerable stage setting would be required.

The sequence of chemical reactions that took place in the Miller-Urey experiment may have been instantiated not only at the dawn of life, and but innumerable times since. In principle then any of those instantiations could be used to exemplify the emergence of organic compounds from inorganic components. All that would be necessary is to ignore all the organic chemicals in the neighborhood; ignore the prevalence of oxygen and nitrogen in the atmosphere; and just concentrate on the reactions due entirely to hydrogen, methane, ammonia, water and electricity. Easier said than done! What distinguishes the experiment from other instances of the sequence is that it constitutes a context where it is manifest that nothing but the four chemicals and a bit of electricity was necessary to get the process started. The experiment thus affords not just an instance, but a telling instance. By exemplifying that those chemicals, and electricity alone suffice to generate organic compounds, the experiment affords genuine insight into an aspect of nature.

An experiment is a dynamic process that unfolds over time. It has a narrative structure (Nersessian 1993), with a well-defined beginning, middle and end. The scientist begins in medias res. She conducts her investigation against a background of established findings and shared assumptions that frame the events and circumscribe their interpretation. The narrative arc

of the Miller-Urey experiment starts with the enclosure of the pure chemicals in a carefully crafted, tightly sealed apparatus, develops through the heating of the water and the emission of an occasional spark, and climaxes when the reactions are complete. The denouement consists in extracting the resulting solution and subjecting it to chromatography to determine its chemical composition.

Experimental results do not speak for themselves. They require interpretation. Their interpretation draws on background knowledge, beliefs about instrumentation, experimental design, the course of events that constitute the experiment, and the outcome. A change in any one of these factors can prompt revisions in others. Against the background of the theory of relativity, we interpret the Michelson-Morley experiment as affording evidence of the non-existence of luminiferous ether; prior to the acceptance of relativity, it afforded evidence that the ether might not exist; long after most physicists accepted the theory of relativity, Michelson took the experiment to demonstrate that his instruments were not sensitive enough to measure ether drift, which he still believed was there to be measured.

These features of experiments are well known. My point in mentioning them is to highlight how distant many scientific experiments and their results are from the natural phenomena they illuminate. The items experimented upon are often artifacts constructed expressly for experimentation. The circumstances in which they are placed are artificial; they are carefully contrived situations, often ones that do not naturally occur but that are designed expressly to exemplify telling features of the phenomena. Experiments are conducted; they do not just happen. They have a narrative structure. They are subject to interpretation and to reinterpretation if background assumptions change. They are repeatable. In short, they are close kin to dramatic enactments. This is not (quite) to say that experiments are works of fiction; but it

is to suggest that the gulf between fact and fiction may be narrower than is typically supposed.

Nevertheless, one might insist, there is a crucial difference. Unlike the 'events' in a novel, the events constituting an experiment actually occur. They are processes that real things really undergo. The Miller-Urey experiment produced real organic compounds from real inorganic chemicals. Moreover, the events in an experiment, unlike those in a play, do not always unfold according to plan. As the Michelson-Morely experiment shows, an experimental result can call into question the assumptions on which it was based; it can provide evidence that reality is not as we take it to be. This independence is crucial to an experiment's epistemic function.

**Thought Experiments**

In standard experiments, scientists simplify, streamline, manipulate and omit, so that the effects of potentially confounding factors are minimized, marginalized, or canceled out. An experiment deliberately departs from nature in order to advance an understanding of nature. Rather than invalidating the experiment, this departure is what enables it to disclose barely detectable, or standardly overshadowed aspects of nature. Thought experiments involve further distancing. They are not actual, and often not even possible, experiments. They are imaginative exercises designed to disclose what would happen if certain conditions were met.

Their reliance on imagination may give hard-nosed epistemologists pause. The imagination is free to entertain any ideas it likes. It is not bound to respect conceptual connections, evidence, laws of nature, or the dictates of common sense. Familiar scientific thought experiments violate all of these. But, as Kant emphasized, freedom is not lawlessness (1993). Freedom consists in being bound by laws we set for ourselves – laws we reflectively endorse as reasonable and rational. As the locus of the free play of ideas, the imagination is not a

realm in which ideas are utterly unconstrained, bouncing off one another like gas molecules in random motion. It is a realm in which the play of ideas is bound by constraints the imaginer sets. Although the constraints are self-imposed and vary from one imaginative context to the next, they are real. The power of thought experiments to illuminate the facts lies in no small measure in the flexible, variable, but nonetheless binding character of the constraints that the imagination imposes on them. By setting such constraints and drawing out the consequences, the imagination serves as a laboratory of the mind, a venue in which hypotheses can be contrived, elaborated, and tested. Moreover, in scientific thought experiments, the constraints, even if tacit, are recognized, shared, and considered appropriate by a scientific community. Even so, how can a thought experiment claim to yield any insight into the facts? Why isn't it simply an exercise in fantasy?

Thought experiments are not essentially private; nor are they particularly mental. Although they are imaginative exercises, they are publicly articulated, discussed, illustrated and disputed. They consist of verbal or pictorial representations. Their claim to be imaginative stems from the fact that like works of fiction they are typically not, and in any case need not be, representations *of* anything real. But the unreality of the objects that ostensibly figure in them does not undermine their function.

In designing a standard experiment, a scientist may begin by performing something like a thought experiment. She runs through the expected course of events in her head or describes it to her research assistants before attempting to implement it in the lab. This suggests that the difference between real experiments and thought experiments lies in the fact that thought experiments sharply truncate the experimental process. They omit the implementation step. But scientists can't do this whenever they please. The question is when is stopping short legitimate?

Sometimes an actual experiment of the sort envisioned *cannot* be carried out. It is

impossible or impracticable. By imagining the experience of a person riding in an elevator in the absence of a gravitational field and at rest in the presence of a gravitational field, Einstein showed the equivalence of gravitational and inertial mass. To actually run the experiment would require placing an unconscious subject in a windowless enclosure, sending him to a region of outer space distant from any significant source of gravity, restoring him to consciousness, and querying him about his experiences. This is morally, practically, and physically unfeasible. Still, the recognition that we cannot do a real experiment does not by itself legitimate the results of stopping short. Sometimes, the infeasibility of an experiment translates into the infeasibility of finding out a particular fact. The reason Einstein's thought experiment is effective is that takes the form of a challenge: Suppose the specified conditions were met. How could a subject tell whether he were in one situation or the other? If our best efforts to identify a way to tell the difference fail, and fail for scientifically principled reasons, we have evidence of the equivalence. Our failure indicates that, if our theories are close to correct, there is no difference to detect.

Sometimes the imaginative rehearsal reveals that an actual experiment *need not* be carried out. The mental run-through itself discloses the relevant information. Without physical implementation, Galileo's thought experiment discredited the Aristotelian contention that the rate at which bodies fall is proportional to their weight. Imagine a composite object consisting of a boulder tethered to a pebble. Being composed of two rocks and some rope, the composite object is heavier than either rock alone. If Aristotle is right, it should fall more quickly than the boulder. But since, according to Aristotle, the pebble falls more slowly than the boulder, once the two are tied together, the pebble should retard the boulder's fall. Hence the rate at which the composite object falls should be between that of the boulder and that of the pebble. The composite object cannot fall both more quickly and more slowly than the boulder, so the

Aristotelian commitments are inconsistent. By exemplifying the inconsistency, Galileo's thought experiment demonstrates that the Aristotelian account cannot be correct.

One might argue that Galileo's thought experiment discredits my analysis. Exemplification, I said, requires instantiation. Real chemical reactions occur in the Miller-Urey experiment. So it is reasonable to think that by exemplifying those reactions, the experiment affords epistemic access to them, enabling us to recognize them and appreciate their significance, not only in the experimental setting but also when they occur elsewhere. In a mere thought experiment, however, nothing actually falls. A thought experiment, not being material, cannot exemplify material properties. This is so. The sequence of ideas that constitutes Galileo's thought experiment does not instantiate material properties of falling bodies. But the *rate* at which bodies fall and the *independence of that rate* from the weight of those bodies are abstract properties. They can be instantiated by material and immaterial sequences alike. So there is no bar to saying that via exemplification thought experiments afford epistemic access to abstract properties that are instantiated in material objects. A thought experiment is a representation – a re-presentation of abstract features, an imaginative re-embodiment of them. We are to imagine – that is mentally, verbally or pictorially present – a situation where the abstract features are realized. In effect, we are to investigate what would happen in a virtual reality where certain constraints are said to hold.[1]

Philosophers sometimes think that we resort to thought experiments only when, for one reason or another, a real experiment cannot be carried out. Perhaps Galileo could not have conducted a real experiment to conclusively demonstrate his point. Maybe he did not have sufficiently accurate timers or a high enough tower from which to run the test. Maybe he did not have the resources to eliminate the effects of air resistance, and so on. Now, however, we could

conduct the experiment. Shouldn't we? Probably not. Rather than concluding that the thought experiment was a second-best strategy resorted to because of circumstances beyond the scientist's control, we should recognize that a real experiment would not have made Galileo's case any more forcefully than his thought experiment did. Indeed, it would simply have muddied the waters. Once we start dropping objects from towers, we face the problem that cancer-ridden wild mice pose for biologists and the emergence of organic chemicals in non-isolated situations pose for chemists. How do we know that unrecognized confounding factors do not explain our finding? By deploying an austere thought experiment where the distance and duration of the fall, the presence or absence of air resistance, and a host of other potential sources of interference are omitted, Galileo blocks such challenges. The thought experiment demonstrates an inconsistency in the Aristotelian position – an inconsistency that would obtain regardless of the conditions under which the experiment was conducted. The thought experiment is preferable to an actual experiment because it is invulnerable to a host of potentially misleading challenges that an actual experiment would face.

Even in the empirical sciences, not every question can or need be answered by direct appeal to observational evidence. Thought experiments are often appropriate where observation is not apt. This leads some to conclude that thought experiments are effective where the issues concern conceptual or theoretical commitments. If so, thought experiments disclose something about our concepts and our theories, not something about the world. Since we can tease out commitments in the armchair, it is no surprise that mere thought experiments are effective for revealing them. But to conclude that thought experiments do not reveal anything about the way the world is would be too hasty. Galileo's thought experiment did not just exemplify an inconsistency in Aristotle's theory. It also showed that any theory that took the rate at which

bodies fall to depend on their weight would be inconsistent. From this it follows that the the rate at which bodies fall is independent of their weight. The thought experiment thus exemplifies a feature of the world, not just of theoretical or conceptual commitments. That feature is, or is a consequence of, a modal fact. The rate at which objects fall is independent of their weight because they could not fall any other way. Thought experiments, it seems, afford epistemic access at least to theoretical, conceptual, and modal matters.

A thought experiment fixes certain parameters (e.g., about the relevant laws of nature and the supposed initial conditions), provides a description of the experimental situation that sets out all and only the features considered relevant, and works out the consequences. Galileo's two rocks are assumed to be made of the same material and to be the same shape, thereby obviating the effects of material and shape on the outcome. They are assumed to be falling through the same medium. Weight is assumed to be additive, and a tethered material object is assumed to be subject to the same laws of nature as untethered ones (Gendler 2010). In effect, the thought experiment invites us to consider would happen if certain conditions, some expressly specified and some tacitly assumed, obtain.

Like literary fictions and ordinary experiments, thought experiments have a narrative structure. We perform thought experiments by imagining a scenario in which something happens – a sequence of events with a beginning, middle and end. Thought experiments can be construed as tightly constrained, highly focused, minimalist fictions, like some of the works of Borges. If the minimalist stories of Borges are genuine fictions, there seems no reason to deny that thought experiments are too.

To understand a thought experiment requires a suspension of disbelief. We grant its (tacit and explicit) assumptions even though we know that they do not – and in some cases cannot –

obtain. Although we know full well that no one is, or arguably could be, endowed with the abilities ascribed to Maxwell's demon, we bracket that knowledge and see whether such a being could defy entropy. Considerable stage setting is sometimes required for it to be clear which beliefs should be suspended and which ones should be retained. This is why scientific thought experiments are embedded in theoretical discussions which fix their parameters. And it is why their implications are subject to dispute.

Like literary works and ordinary experiments, thought experiments require interpretation. Sometimes interpretations diverge. The Einstein-Podolsky-Rosen experiment is a case in point. Very roughly, the scenario is this: Two particles interact, then fly off in opposite directions. Once they are separated, the measurement of one should have no effect on the state of the other. But if we measure, say, the position of one and apply the Schrödinger equation, we can determine that the other also has a definite position. This seems to violate the uncertainty principle: the unexamined particle should have no definite position. What does the thought experiment show? The answer is not at all clear. Apparently, even the authors disagreed. Podolsky thought it demonstrated that quantum mechanics is incomplete, while Einstein thought it showed either that quantum mechanics is incomplete or that states of spatially separated objects are not independent of each other (Bokulich 2001).

To recap: a thought experiment is an imaginative exercise designed to investigate what would happen if certain conditions were satisfied. Conducting it requires a suspension of belief, in that the conditions imagined are not realized in fact, and may be inconsistent with conditions we know to obtain in fact. It requires a suspension of disbelief, in that it asks us to entertain scenarios that we know do not and often could not obtain. It depends on background assumptions about what commitments are to be retained, what commitments are to be relaxed,

and what commitments are to be abandoned in entertaining the imaginative scenario. Whether the constraints are tacit or explicit, in conducting the thought experiment the epistemic agent is bound by them. A thought experiment has a narrative structure, with a beginning, middle and end. It is subject to interpretation, and to reinterpretation if the background assumptions change. Schrödinger's cat, originally introduced to criticize the Copenhagen interpretation, now appears in every interpretation of quantum mechanics, each offering a different account of the poor beast's state. Finally, we saw that thought experiments are valuable in investigating what is not open to direct empirical inspection – conceptual or theoretical commitments and their consequences, as well as modal properties, and, I would add, relations of cotenability on non-cotenability and so on.

**Fictions**

The thought experiments I have mentioned so far have been drawn from the physical sciences. But thought experiments are ubiquitous in philosophy as well. Like scientific thought experiments, those in philosophy illuminate factors that are not accessible to direct inspection. But because of a difference in subject matter, the range of factors illuminated by philosophical thought experiments is broader. Familiar philosophical thought experiments afford insights into normative properties (trolley problems, the experience machine), introspectively available properties (Mary, brains in a vat), and metaphysical properties (fission and fusion in personal identity, the ship of Theseus). No more than conceptual, theoretical, and modal properties, are these open to direct empirical inspection.

Like scientific thought experiments, philosophical ones tend to be fairly austere. But although they are typically performed in the context of theorizing, philosophical thought

experiments often are not so tightly framed by theoretical constraints as scientific thought experiments tend to be. Thus there is more controversy over what we should conclude from them. (Confession: I have no idea what the Chinese Room shows.) Moreover, many philosophical thought experiments are relatively autonomous. We can fruitfully entertain them against the background of multiple sets of philosophical assumptions – indeed without even being aware that we are making philosophical assumptions. The trolley problem was introduced to disclose a consequence of the doctrine of double effect. It now has a life of its own (and numerous children and grandchildren).

If an austere thought experiment can afford epistemic access to a range of properties, and can do so in a context that is not tightly beholden to a particular theory, there seems to be no reason to deny that a more extensive thought experiment can do the same. This opens the way to construing works of literary fiction as extended, elaborate thought experiments. They afford epistemic access to aspects of the world that are normally inaccessible – in particular, to the normative, psychological and metaphysical aspects that philosophical thought experiments concern.

Again one might worry about the capacity of exemplification to account for this. A work of fiction, not being alive, cannot instantiate psychological or moral properties. If it cannot instantiate them, it cannot exemplify them. This is true. But it can instantiate and exemplify abstract properties that are concretized in human agents. Suppose Meg instantiates a pattern of psychological features – a network of beliefs, desires, and preferences, for example. Although the specific elements of her network are psychological, the pattern is abstract. In principle, it can be instantiated by something other than psychological elements. A fiction writer might create a scenario where that pattern is instantiated and exemplified via a sequence of descriptions. In

effect, she takes a pattern that is embodied in fact, abstracts it, and re-embodies it in fiction. (Or, more likely, she abstracts individual elements instantiated in fact, finds or devises an appropriate pattern, and embodies that pattern in fiction.) Strictly, in the fictional setting it is not a pattern *of* psychological features. But it is a pattern that is, or that may be, instantiated by psychological features. So it affords epistemic access to a pattern that we may find ourselves or our fellows instantiating.

Like an experiment, a work of fiction selects and isolates, contriving situations and manipulating circumstances so that patterns and properties stand out. It may frame or isolate mundane features of experience so that their significance is evident. It may defamiliarize the commonplace, making us aware of how remarkable normal behavior can be.

Jane Austen evidently agreed. She wrote to her niece, 'Three or four families in a country village is the very thing to work on' (1814). Because the relations among the members of three or four suitably characterized families are sufficiently complicated, and the demands of village life sufficiently mundane, her stories can exemplify something worth noting about ordinary life and the development of moral personality. In limiting herself to three or four fictional families, Austen devises a tightly controlled thought experiment. Restricting the factors that impinge on her protagonists enables her to elaborate the effects of those that remain. 'Though more simplified and structured than actual cases, [fictions] are much richer in detail – about motives, feelings, circumstances, social relations, and interconnected personality traits' (Carroll 2002, p. 19).

Wouldn't it be cognitively preferable to study three or four real families in a real country village? Probably not, if we want to glean the insights that Austen's novels afford. The problem is akin to the one that we saw with actually running Galileo's thought experiment. Real families,

however isolated, are affected by too many forces for the social and moral trajectories exhibited by Austen's characters to stand out. Too many other factors impinge on them; too many descriptions are available for characterizing their lives. Any such sociological study would be vulnerable to the worry that unexamined factors played a non-negligible role in the interactions studied, that other forces were significant. Moreover, such an empirical study would yield no direct access to motives or feelings and at best indirect access to personality traits. It would be restricted to patterns of interaction investigators happened to find. A given village might fail to contain inhabitants with the personality traits of Mr. Darcy and Elizabeth Bennet, so the interaction Austen was bent on studying would not be found. Unlike a sociologist, Austen could construct the personalities whose interactions she wanted to investigate, and put her protagonists in situations where their telling features reveal themselves.

In *The Nicomachean Ethics,* Aristotle suggests that we should call no man happy until he is dead. Initially this seems implausible. As is well known, part of the problem is that what Aristotle means by 'happy' is not what we mean. 'Flourishing' would be a better term. But 'call no man flourishing until he is dead', even if less implausible, still seems extreme. Surely, one wants to object, we can easily discern that some of our fellows are currently flourishing. Aristotle defends his idea by contending that severe enough reversals of fortune late in life would justify the conclusion that a man's life had not been a happy (or flourishing) one. Maybe so. But even if someone suffered serious misfortunes late in life, it is tempting to say, 'Well he was happy (or flourishing) up until then'. Aristotle's own example of Priam, the elderly king of Troy, is vulnerable to this objection. Priam was evidently thriving until the Trojan War. His life ended in misfortune; but throughout most of it, he seems to have flourished. This is the argument that students typically give against Aristotle, and if one just reads Aristotle, it does not seem

unreasonable.

*Oedipus Rex* can be read as a thought experiment that vindicates Aristotle's claim. For most of his life Oedipus, like Priam, seemed blessed with the gifts of fortune; and as far as anyone could tell, he lived a life of Aristotelian virtue. Evidently, he flourished and deserved to flourish. Then Thebes suffered a plague and the oracle blamed it on him. Oedipus discovered that, unbeknownst to himself, he had killed his father and married his mother. This discovery did not just doom his future happiness, as the defeat of the Trojans doomed Priam's. It discredited his past happiness. He had, through no fault of his own, been living a lie. We might exonerate him for the wrongs he had done, since he acted out of nonculpable ignorance. But even if he was blameless, his relations to himself, to his wife/mother, to his children/siblings, to the citizens of Thebes who suffered for his iniquities, and to his own past, were forever changed, and would henceforth be tinged with revulsion. It turns out that he had not been flourishing during the early years, even though he and everyone else thought that he was.

*Oedipus Rex* is a work of fiction that advances our understanding not only of Aristotle's ethics, but also of the human predicament. It underscores the limits on human knowledge and the vulnerabilities that stem from those limits, the value of knowing oneself and one's situation, and the limits on the human capacity to do so. It does not constitute a proof that Aristotle is right, but it poses a challenge: you should be wary of calling a man happy during his lifetime unless you are sure his situation is not like Oedipus's. It is hard to see how the challenge can be met.

Kant maintains that not only is it impossible to know whether someone else has acted morally, it is impossible to know whether you yourself have done so (1993). For any action that accords with the categorical imperative, there is always an available self-interested maxim that

might have been the real motive. So it is impossible to glean unequivocal empirical evidence for moral action. But even if we can never discern that someone has behaved morally (that is, acted on account of respect for the moral law), an author can effectively stipulate it. In *A Tale of Two Cities* Dickens portrays Sidney Carton's self-sacrifice as stemming from purely moral motives. An author can portray a situation, develop a character, and convey his thoughts, feelings, and reasons for acting, thereby blocking the explanation from self-interest. He can, that is exemplify a pattern and demonstrate that were that pattern to be instantiated, it would be an instance of acting morally.

If Kant is right about the impossibility finding unequivocal instances of acting morally, how do we learn what a moral action is? Even if the categorical imperative is a deliverance of reason, specific applications need to be learned. Non-kantians face a similar problem. Everyone recognizes that it is difficult in practice to distinguish acting morally from acting out of self-interest. If I am right, fiction can play major role in moral education. By exemplifying genuinely moral patterns of deliberation and action, it can teach us something that is at least hard, and may be impossible, to reliably discern in real life interactions.

People have inner lives replete with motivations, perceptions, emotions and thoughts. Because a variety of combinations of psychological elements might yield the same outward behavior, it is impossible to uniquely determine the underlying psychological states from observations of overt behavior alone. Moreover, people often misunderstand themselves, and often have reason to mask what they feel. So even if we ask them and they tell us, we ought not be confident that we know how things are with them. This raises the question: how do we learn (or what leads us to think) that other people have inner lives that are quite unlike ours – not only unlike what we actually feel about things, but quite unlike what we would feel if we were in their

place? It is one thing to be able to imagine oneself having had someone else's experiences. That is hard enough. But we can do things that require yet more imaginative dexterity. Even if Jim recognizes that had he been treated as Jane was, he would be bitter, he may also understand why she is willing to let bygones be bygones. This is an amazing cognitive accomplishment. I believe that fiction plays a major role in equipping us for it. In reading a work of fiction we take up a point of view and try it on for size. In effect, we experiment with the perspective and see how things look from there. Many works portray the world through a protagonist's eyes, conveying her experiences, feelings, and thoughts. They disclose the limitations of her perspective. Some do more. A work may afford multiple perspectives on the same series of events, disclosing the resources and limitations of each. In effect, each filters events through a different sieve.

Philosophers as well as non-philosophers have a tendency to take the fruits of introspection at face value, or at least to grant them a higher epistemic status than outsiders' opinions about a subject's state of mind. A work like *Lolita,* written from the perspective of an utterly unreliable narrator, affords insight into the limits of introspection. A character's perspective can be so skewed or benighted that he is simply wrong about the central events of his life. Such a work can be construed as a thought experiment that undermines the conviction that a person's access to his own motives, beliefs, and other attitudes always affords better evidence than the evidence that one's words and actions afford to others. Perhaps there is privileged access in the sense that each of us knows herself *in a way* that she knows no one else. But self-deceptive fictional characters undermine the conviction that we always know ourselves *better* than others know us.

Cavell maintains that the problem of other minds is not, or not only, the problem of

ascertaining whether an entity has a mind, but the problem of figuring out what is on someone's mind. Even if I am confident that an individual has beliefs, desires, preferences, and feelings, I am woefully underequipped to identify those fine-grained mental states. The problem is not just a problem about other minds, though. I may be equally underequipped to know what my mental states are. Do I really expect I'll get the paper done, or am I deceiving myself into taking a hope for an expectation? Do I really desire the promotion or do I just think I desire it because I know that it is the sort of thing that people in my position are supposed to desire? As Cavell (1987) reads Shakespeare's tragedies, they afford evidence of the uncertainty of mental state ascriptions. For the same sorts of reasons that Lear cannot recognize Cordelia's devotion, that Othello cannot recognize Iago's malevolence and Desdemona's fidelity, that Hamlet cannot trust his judgment, we cannot be sure of the mental states that we ascribe.

A work of literature can function as something akin to an impossibility proof – a thought experiment that exemplifies the inadequacy of its grounding assumptions. Davenport (1983) reads *Middlemarch* as a thought experiment about marriage. Both Dorothea Brooke and Dr. Lydgate are in deeply unhappy marriages. Because in the world of the novel divorce is unthinkable (and unthought of), they are destined to serve life sentences for their unwise choices of mates. By exemplifying the intractability of the problem they face, the novel affords reason to think that divorce, or something like it, should be an option. Davenport considers *Middlemarch* flawed because it does not allow for the possibility of divorce. I disagree. I consider it a powerful thought experiment that reveals the consequences of institutional structures that do not allow for divorce.

Metaphysical thought experiments are often science fictional. Some are so austere that in their philosophical settings we do not know what to think. Literary and cinematic fictions help

us out. What should we make of Putnam's brains in a vat? *The Matrix* supplies an answer. What would a computer that passed the Turing test be like? His name is Hal. Could beings without any inner lives actually be indistinguishable from us? The way to settle such matters (even tentatively and revisably) is to design a scenario in which the consequences of such hypotheses play out. Write a story about the love lives of zombies, or about the lives of zombies incapable of love. We may find that our off the cuff intuitions do not stand up under elaboration, or that the consequences of our assumptions are quite different from what the austere philosophical thought experiments led us to suppose.

Despite what I have said, the idea that fictions function as thought experiments that afford insight into human experience may seem a stretch. Let me give an example. One of the mysteries of the recent Penn State pedophilia scandal is how Joe Paterno, the longtime football coach whom many held to be an bastion of integrity, could have turned a blind eye to the actions of his assistant. Sportswriter Thomas Boswell ventures the following answer:

> Everybody has weak spots in their character, fault lines where the right earthquake at the wrong time can lead to personal catastrophe. Most of us are fortunate that our worst experience doesn't hit us with its biggest jolt in exactly the areas where our flaws or poor judgment or vanity is most dangerously in play. It's part good luck if we don't disgrace ourselves. But when it does happen, as appears to be the case with Joe Paterno, that's when we witness personal disasters that seem so painful and, in the context of a well-lived life, so unfair that we feel a deep sadness even as we simultaneously realize that the person at the center of the storm can never avoid full accountability. . . . Forces collide, conspire, confuse, and an icon of integrity fails to act, fails to see. (Boswell 2011).

If this sounds familiar to those who do not read the sports pages, it is because the passage is a

precis of *Oedipus Rex* as filtered through Aristotle's *Poetics*.[2] The great man, beset by hubris, does terrible things and is brought down by his tragic flaw. One can quarrel with Aristotle's reading of Sophocles or with Boswell's implicit endorsement of that reading. One can doubt that Paterno was the man of integrity he was alleged to be. One can even think that ignoring rampant pedophilia is worse than inadvertent patricide and incest among consenting adults. Still, *Oedipus Rex* affords a template for understanding Paterno. Having seen the pattern in fiction, we are in a position to entertain the possibility that it explains what happened in fact. As David Lewis notes,

> We who have lived in the world for a while have plenty of evidence, but we may not have learned as much from it as we could have done. This evidence bears on a certain proposition. If only that proposition is formulated, straightway it will be apparent that we have very good evidence for it. . . . If we are given a fiction such that the proposition is obviously true in it, we are led to ask: and is it also true *simpliciter*? And sometimes, when we have plenty of unappreciated evidence, to ask the question is to know the answer. (Lewis 1983, p. 279.)

Experiments yield evidence, not proof. And evidence is sometimes misleading. The status of an item as an experiment or thought experiment does not hinge on its being successful in advancing understanding. I do not claim that every work of fiction succeeds. Nor do I claim that every experiment or thought experiment does. Some are muddled or confused. Some overlook real possibilities or fail to control for important variables. Some replicate what is already known or widely accepted. Some are trivial. Some thought experiments and literary works may even be the equivalent of high school science experiments, exemplifying what is already understood in an effort to show how such symbols advance understanding. Any finding

must be tested by its fit with what we already have reason to believe. Galileo's thought experiment revealed an inconsistency in Aristotle's theory. It is, or is very close to, a crucial experiment. We need nothing further to show that Aristotle's theory is false. But most experiments and thought experiments, and most works of literature, work within a context of background assumptions. If the assumptions are incorrect or incomplete, an experiment or thought experiment may inherit and reinforce their inadequacies. If they are (close enough to) correct and complete, the experimental result is prima facie informative.

Still, there are reasons why we might resist identifying works of fiction with thought experiments. One is what Carroll (p. 4) calls the argument from banality, the contention that the knowledge imparted by fictions amounts to little more than truisms.[3] The guiding idea seems to be that if fiction has an epistemic function, it is to impart ethical truths like the morals of Aesop's fables. These truths are inferred inductively or deductively from stories. Since such truths are banal, they are, for the most part at least, epistemically inert. But the patterns and features that works of fiction exemplify are far from truisms. Because exemplars display, rather than merely state, they can be exceedingly fine-grained. Hitchcock's *North by Northwest* exemplifies delicate nuances in the texture of fear – a virtual continuum from trepidation to terror. There is nothing banal about learning to recognize subtle differences and project them properly onto members of the classes the exemplars typify. Nor is it likely that we will be able to capture in a pithy proposition just what such a fiction discloses.

That a work is a rich source of insights is not a reason to doubt that it advances understanding. But it may be a reason to doubt that the work is a thought experiment. Stereotypical thought experiments tend to be austere. Although they require interpretation, their interpretations are supposed to be univocal, at least until the relevant background assumptions

change. But univocality is not a virtue in literary or dramatic fictions. That *Henry V* can be interpreted as pro-war and anti-war is not a defect in the play. Literary works are semantically dense and replete (Goodman 1968). Works of fiction are apt to bear multiple correct interpretations. So in this respect they differ from austere, univocal thought experiments. Perhaps this is a reason to deny that works of fiction are thought experiments; perhaps it is a reason to say that only under an interpretation is a work of fiction a thought experiment; perhaps it is a reason to think that some thought experiments are more austere than others.

I favor the last option. Although stereotypical thought experiments are austere, there is a continuum of cases from Maxwell's demon and trolley problems through the myth of the cave and *Emile* to 'didactic fictions' like *Animal Farm* and *Uncle Tom's Cabin,* to *Middlemarch* and *Oedipus Rex.* I doubt that there is a sharp boundary between thought experiments, strictly so called, and works of fiction. But demarcating the boundary is not so important. Whether or not we call works of fiction thought experiments, I have urged that fictions, thought experiments, and standard experiments function in much the same way. By distancing themselves from the facts, by resorting to artifices, by bracketing a variety of things known to be true, all three exemplify features they share with the facts. Since these features may be difficult or impossible to discern in our everyday encounters with things, fictions, thought experiments and standard experiments advance our understanding of the worlds and of ourselves.[4]

## References


Aristotle. 1941. 'Nicomachean Ethics,' in *The Basic Works of Aristotle,* ed. Richard McKeon. New York: Random House, pp. 935-1026.

Aristotle. 1941. 'Poetics,' in *The Basic Works of Aristotle,* ed. Richard McKeon. New York: Random House, pp. 1455-1487.

Austen, Jane. 1814. Letter to her niece, Anna Austin Lefroy, September 9, 1814, in *Letters of Jane Austen* Bradbourne Edition, www.pemberley.com/janeinfo/brblets.html. Consulted May 4, 2005.

Ball, Philip. 2005. *Elegant Solutions.* Cambridge, UK: Royal Society of Chemistry.

Bennett, Jonathan. 1974. 'The Conscience of Huckleberry Finn', *Philosophy* 69: 123-134.

Bokulich, Alisa. 2001. 'Rethinking Thought Experiments' *Perspectives on Science* 9: 285-307.

Boswell, Thomas. 2011. 'Penn State Coach Joe Paterno Reaches a Sad Conclusion' *Washington Post,* November 9.

Carroll, Noel. 2002. 'The Wheel of Virtue: Art, Literature, and Moral Knowledge,' *The Journal of Aesthetics and Art Criticism.* 60: 3-26.

Cavell, Stanley. 1987. *Disowning Knowledge in Six Plays of Shakespeare.* Cambridge UK: Cambridge University Press.

Davenport, Edward. 1983. 'Literature as Thought Experiment (On Aiding and Abetting the Muse). *Philosophy of Social Science* 13: 279-306.

Elgin, Catherine. 1996. *Considered Judgment.* Princeton: Princeton University Press.

Fricker, Miranda. 2007. *Epistemic Injustice.* New York: Oxford University Press.

Gendler, Tamar. 2010. 'Galileo and the Indispensability of Scientific Thought Experiment' in her *Intuition, Imagination and Scientific Methodology.* Oxford: Oxford University Press, pp.

21-41.

Goodman, Nelson. 1968. *Languages of Art*. Indianapolis: Hackett.

James, William. 1983. *Principles of Psychology.* Cambridge, MA.

Kant, Immanuel. 1993. *Grounding of the Metaphysics of Morals.* Indianapolis: Hackett.

Lewis, David. 1983. 'Truth in Fiction: Postscript' in his *Philosophical Papers* vol 1. Oxford: Oxford University Press, pp. 276-280.

Nersessian, Nancy. 1993. 'In the Theoretician's Laboratory: Thought Experimenting as Mental Modeling,' *PSA 1992* Volume 2. Philosophy of Science Association: 291-301.

Snyder, Martin. 2012. 'Teaching Jo Pa,' *Academe* 98: 55.

[1] This is consonant with Platonism, but does not require it. Perhaps abstract properties and patterns exist only if instantiated, but instantiations, whether material or virtual, and can be created or can naturally emerge.

[2] I originally found reference to is column in Martin Snyder's 'Teaching Jo Pa', *Academe* 98, p. 55.

[3] Carroll discusses but does not accept this argument.

[4] I am grateful to Jonathan Adler, Geordie McComb and Amelie Rorty for helpful comments on earlier drafts of this paper.