Most philosophers, it seems, like things neat and tidy; many hold on to the ideal that we should aim at revealing the deep structures of reality by performing clean cuts that – to use Plato’s oft-quoted metaphor — »carve nature at its joints.«¹ They are not likely to accept the very idea that reality might be unstructured, messy, or chaotic, nothing but a »blooming, buzzing confusion,«² arbitrarily composed of a vast range of phenomena of the most different kinds. According to this image, the unstructured manifold that presents itself to our senses is overwhelming only for the untrained eye of the novice – while the sage, who knows what to look for, can discern the deeper, underlying structure of reality. In order to do so, one has to have attuned oneself – or better: one’s perceptual apparatus – to the relevant aspects of our environment, one has to have figured out which properties indicate or instantiate the deep, underlying structure of reality. The natural sciences, it is often suggested, provide for privileged ways to individuate the properties that should be considered salient – or »élite« – properties,³ and scientific experiments – including thought experiments – are their most powerful means to do so.

The philosophers’ fondness of things being neat and tidy also exists in the philosophy of literature. This tendency, in my point of view, sometimes constitutes a serious obstacle for finding the most obvious solutions to the problems that are at the center of debate (and which, incidentally, are often presented in the form of a paradox): It is not always clear (to me) that philosophers who work in the field are willing to appreciate that the very term »literature« does not carve nature – in this case the nature of artifacts or socially constructed phenomena – at its joints. The term is used for a huge variety of phenomena – ranging from historical novels to fairy tales, from autobiography to science fiction, from Greek tragedy to Dadaist poetry. The very conception of literature has changed essentially over the centuries and varies substantially from one culture to another. Moreover, the differences are not only to be located in the text: As Peter Kivy has recently pointed out, there are also great differences among readers (Kivy 2011, 32ff), who approach works of literature with very different interests and expectations. Thus, we should not expect to be able to find a short, clear-cut, and homogeneous answer to the question concerning the nature and value of »literature«; we should rather be prepared to end up with a long, detailed, and multifarious account.

¹ Cf. Plato Phaedrus [265e].
³ The expression »élite properties« was coined by David Lewis (1984); for an interesting discussion, cf. Elgin (1995).
A quick look at the current debate within philosophy of literature shows that the complexity of the phenomenon of literature is often acknowledged – with no obvious consequences: Most contributions continue to treat literature as a uniform phenomenon. This attitude becomes manifest not only in the generality of the formulations of the respective theses but also in the works that are discussed: Most contributions focus on a very specific form of literature, typically on fictional narratives, and in particular on novels or short stories\(^4\), but the results are presented as if they would hold for literature \textit{tout court}.

The debate concerning the cognitive value of literature is a paradigmatic example of this tendency. Anti-cognitivists typically deny that fictional literature could ever be of cognitive value or, more precisely, that any work of literary fiction could impart knowledge to the reader on the basis of its distinctively \textit{literary} features. The cognitivist replies take a similar line though. even the most interesting contributions aim at individuating a \textit{single} mechanism in virtue of which works of fictional narrative can impart forms of propositional knowledge or forms of understanding to the reader – and that is particularly apt for some but can be less suited or even inadequate for other (kinds of) literary works.

The suggestions, for example, that works of literary fiction prompt an exercise of imagination in a game of make-believe (Walton 1990) or that they deepen our understanding through presentification [\textit{Vergegenwärtigung}] (Gabriel 1991, 1997) focus – very much like the (related) views that literature provides forms of non-propositional knowledge (Schildknecht 2007) or an understanding of the perspective of others (Feagin 1996, 2007) – on the fact that many works of literature contain detailed descriptions of counter-factual scenarios. The suggestions that literature can refine our moral perception (Nussbaum 1990) or enrich us not at the level of knowledge but of acknowledgment (Gibson 2003, 2007) focus very much on the plenty of detail and the \textit{human face} of the situations described while the proposal that it can reinforce our conceptual capacities and that of making inferences (Huemer 2007b, 2007a) considers mainly formal aspects of texts as relevant. The list of examples – which could easily be extended – shows that most proposals raise a point that can explain very well how \textit{some} works of literature can contribute to our knowledge or understanding but cannot do justice to others. What about the approach that is at the focus of the present volume: Can the view that literature is a thought experiment give a more comprehensive explanation of the cognitive value of literature?

\textbf{Some Observations on Thought Experiments}

\(^4\) A short look at the most discussed examples in the debate can corroborate the point: Prominent scholars have contented themselves to ponder about the ontological status of Sherlock Holmes, our emotions towards Anna Karenina, or the cognitive value of Jane Austen novels, often even without discussing the literary works in questions.
In recent years, several philosophers have highlighted the similarities that hold between works of literary fiction and thought experiments (cf. for example, Carroll 2002; Elgin 2007, 2014; Davies 2007; Green 2017): both literary fiction and thought experiments involve essentially our power to imagine counterfactual scenarios, both develop fictional narratives, and both can serve to illustrate a given hypothesis. There is, thus, a strong intuitive force in this analogy, which definitely raises its plausibility – but it might also raise our (Wittgensteinian) suspicions: No analogy is perfect – the stronger the intuitive force of an analogy, the stronger is its power to mislead us. In his *Philosophical Investigations*, Wittgenstein has often warned of this danger, most explicitly in § 115, where he explains the errors of his early philosophy by stating: »A *picture* held us captive.« (Wittgenstein 2009, § 115) In order to evaluate the proposal, it will be useful to briefly highlight the analogies but also the differences that subsist between works of literary fiction and thought experiments. I will do so by focusing on two points that are often discussed in the debate.

(i) **Controlled scenarios**

Both works of literary fiction and thought experiments develop carefully controlled scenarios where all irrelevant parameters are excluded. Moreover, in both cases the scenarios described are not interesting in themselves but rather in their being exemplary for a great number of other cases. Thus, both works of fiction and thought experiments point beyond themselves and aim at a more general level of understanding. They share these features with laboratory experiments although we can note revealing differences: In laboratory experiments it is important to control all causal forces within the experiment because impurities could unfold an unexpected causal efficacy that is not tracked by the experimenter, which can result in a systematic distortion of the results. In thought experiments, on the other hand, where it is much easier to exclude irrelevant factors, the exclusion of irrelevant details is a maxim of efficacy: Our minds are finite and so is our power to imagine counter-factual scenarios. The more complex a thought experiment becomes and the more unnecessary details it contains, the more difficult it will be for the audience to grasp the relevant point. The focus on what is strictly relevant, thus, is a standard dictated by the limits of our powers of imagination – and its tendency to easily go astray.

Many advocates of the analogy between literary fiction and thought experiments have pointed out that also works of literary fiction develop controlled scenarios. Typically, they focus on one or a handful of protagonists and very often they situate them in specific settings to »see what happens«, as it were. Nonetheless, we can note that the sobriety of (laboratory and thought) experiments stands

---

5 Catherine Elgin likes to remind of Jane Austen’s statement that »[T]hree or four families in a country village is the very thing to work on.« (in a letter to her niece in September 1814, quoted in Elgin (2014, 232f))
in contrast to the opulence of fictional literature that attracts with its plenty of (seemingly) «unnecessary details» (cf. Rorty 1991, 80f) and the complexity of narrative. While thought experiments want to prevent the reader’s imagination from going astray, literary fictional narratives seem to invite for it. The contrast between the sobriety of thought experiments and the opulence of literary works is also manifest in their aesthetic form: While thought experiments are typically presented in sober and reduced style that avoids ambiguity and ornament, literary works of art are open to competing interpretations and aim at attracting the reader’s attention on the basis of their aesthetic form.\(^6\)

(ii) **Imagination and Arguments**

Another similarity between thought experiments and works of literary fiction consists in the fact that both invite for an exercise of the reader’s power of imagination. They get the reader involved. Moreover, if they enrich the reader’s understanding; they arguably do so not by transmitting bits of true information that is passively absorbed by the reader but rather by triggering a series of focused reflections. In a way, both thought experiments and works of literary fiction spur the reader’s imagination and often invite her to draw certain conclusions, but she has to make the relevant inferences and come to the respective conclusions herself.

There are substantial differences in the way in which the reader is guided to the proposed conclusion in thought experiments and in works of literary fiction. In philosophy of science there is some debate on how we can learn from thought experiments. While some philosophers, most notoriously James Brown, have conceived them as autonomous sources of knowledge, as means of grasping abstract principles (cf. Brown 2011), others have argued that thought experiments are arguments »disguised in a vivid pictorial or narrative form.« (Norton 2004, 45) There is wide agreement, however, on the fact that thought experiments serve to illustrate or justify (or show wrong) a hypothesis. The reader, thus, is invited to activate her imaginative and cognitive powers, but she is guided on a short leash to a conclusion that is often stated explicitly.

Literary works of art, on the other hand, are typically open to a manifold of competing interpretations. Very often they do point towards a more general »message« – but there is likely not one single »message« that could be extracted from the text; any attempt to make explicit in a single formulation »the« hypothesis that is argued for by a literary text has to appear hopelessly reductive.

In short, thought experiments pursue their cognitive goals in a more direct and straightforward way than literary works of art. In consequence, we can criticize thought experiments for leading to

---

\(^6\) David Egan insists on this point in his article »Literature and Thought Experiment«, where he warns against the dangers of an instrumentalization of literature. I fully share Egan’s point that in order to get a more adequate conception of literature we need to pay attention to the »distinctively literary features of works of literary fiction.« (Egan 2016, 149)
wrong conclusions, for being incorrect or misleading while any attempt to show that a literary work of art – even one that clearly aims at broadening the readers’ cognitive horizon – is wrong would have to be considered inappropriate. One can, of course, suggest that the choice of a given novel’s topic is tendentious, that certain descriptions are superficial or incredible, or that they unveil an implicit bias of some sort, but literary critique typically focuses on the aesthetic features of the work not at their veracity.  

*Dürrenmatt’s The Physicists*

The short discussion of the preceding section illustrates that the analogy between thought experiments and literary works of art can be very stimulating and helpful when it comes to account for the cognitive value of some (but not all) works of literary fiction – but also that it has its limits. In the present section I will try to apply this analogy to a work that seems to be – at least at first glance – particularly apt to be considered a thought experiment: Friedrich Dürrenmatt’s play *The Physicists*.

The play opens with a murder: In an idyllic Swiss sanatorium for the mentally ill run by a famed psychiatrist Fräulein von Zahnd, a patient has killed his nurse. The police come to examine the scene. Chief inspector Richard Voß cannot hide his frustration for the case seems clear and there are not many consequences to be expected as it will not be possible to hold the patient – who believes to be Albert Einstein – responsible for his action. Soon we learn that a similar event has taken place only three months earlier, when another patient – who believes to be Sir Albert Newton – had killed his nurse. The two patients are in a special part of the sanatorium that they share with a third physicist, Moebius, who claims that the biblical King Salomon visits him regularly in his room to dictate him a new physical theory, including the »universal formula« and a »System of All Possible Inventions.«

The nurse who cares for Moebius, however, is not convinced that he is mentally ill; she thinks him a genius and has arranged for him to leave the sanatorium and to move in with her, where she could provide all the support needed for him to go on to work on his theory. Moebius, who seems to reciprocate her love, sees no other way than killing her – for he has a secret to protect: He is, indeed, only faking his insanity and works in the sanatorium on a ground-breaking new physical theory, which, however, would make the construction of most powerful weapons of mass destruction possible. Scared of the potential harm his research could bring about, he has decided to sacrifice his own life, to fake mental illness, and to retire to the sanatorium so as to keep his theory secret and protect the rest of humanity.

---

7 This point was already suggested by Rudolf Carnap, who argued that »[l]yrical poets … do not try to refute in their poem the statements in a poem by some other lyrical poet; for they know they are in the domain of art and not in the domain of theory.« (Carnap 1966, 79f)
We soon find out that both Einstein and Newton had killed their nurses for similar reasons: Like Moebius they were faking their insanity; in reality they are physicists who work for the American and the Russian secret services, respectively. Both sides had figured out that Moebius was working on something important and so decided to infiltrate the sanatorium. Both agents now try to convince Moebius to work for their countries, but it is Moebius, who succeeds in convincing them that the results of his research are too dangerous; sharing it with the wrong persons could easily result in the extinction of the human kind. Therefore, the three decide to sacrifice their own lives and remain in the sanatorium.

The story ends with a final twist when, the moment the three physicists have resigned to their fate, they have to learn that Fräulein von Zahnd, the director of the sanatorium, had known all the time that the three were only faking their insanity. Claiming to have acted on King Salomon’s direct order, she had secretly copied Moebius’ work and used the »System of All Possible Inventions« to build up an empire of companies that has made her become one of the most powerful persons. In fact, soon she will make a final step to gain power over the whole world and subdue the entire human kind. The three physicists understand the vanity of their decision and fall into resignation.

Dürrenmatt’s play, it seems to me, wants to be considered at least prima facie as a thought experiment. The whole setup – there is only a handful of protagonists and the whole story unfolds in a small room and respects the unity of time and place – is precisely circumscribed and apt to exclude irrelevant parameters. Moreover, in the paratext of the play, Dürrenmatt has formulated »21 Points to the Physicists«, which can give a guide to the interpretation of the work. Among these we find:

16. The content of physics is the concern of physicists, its effect the concern of all men.
17. What concerns everyone can only be resolved by everyone.
18. Each attempt of an individual to resolve for himself what is the concern of everyone is doomed to fail. (Dürrenmatt 1973, appendix)

Dürrenmatt’s Physicists is, thus, one of the rare examples of a work of literary fiction where the author adds – at least in the paratext – clear-cut and explicitly formulated theses. From here it seems a small step to suggest that the play is best understood as a thought experiment that illustrates or argues for (at least some of) the theses contained in the list. In particular, it could be argued that the play exemplifies cases where scientists aim to make decisions that concern the whole human kind without opening themselves to discuss or explain their decisions to others and it vividly illustrates, with the counter-factual scenario that is developed, the dangers of scientifically informed technological progress. To be sure, the actual scenario and the twists in the play are highly artificial and have

---

8 The 21 points were written in 1962, the year of the premiere of the play in Zurich and are now printed with the text.
little to do with the life-world of the audience, but this, one could argue, even underlines the impression that we are dealing with a (thought) experiment that is conducted under controlled »laboratory conditions«. In sum, there are reasons to suggest that Dürrenmatt’s *Physicists* is one of the prototypical examples of a literary work of art that can contribute to our understanding by presenting a thought experiment.

At this point we should remind ourselves of Wittgenstein’s warning that pictures that are too powerful can become dangerous. In fact, I think that the interpretation of Dürrenmatt’s play as an enacted thought experiment risks being overly reductive and overlooking aspects that are most relevant in the work. These aspects become more salient if we look at the first five of Dürrenmatt 21 points:

1. I don’t start with a thesis but with a story.
2. If you start with a story, you must think it to its conclusion.
3. A story has been thought to its conclusion when it has taken its worst possible turn.
4. The worst possible turn is not foreseeable. It occurs by accident.
5. The art of the playwright consists in employing, to the most effective degree possible, accident within the action. (Dürrenmatt 1973, appendix)\(^9\)

With his first point, Dürrenmatt suggests that he began with the composition of this play not with a thesis – a thesis that could be the conclusion of an argument if thought experiments in fiction had the structure of arguments – but with a story in his mind. It is as if he wanted to remind us that it is not the (abstract) thesis that counts but the narrative – and a narrative has a plot and leads to an end; one cannot stop narrating at some arbitrary point because otherwise the structure of the narrative would collapse.

Moreover, it seems to me that with his famous third point, Dürrenmatt distances himself even more from the notion of thought experiment. He makes very clear that he does not mean to situate the protagonists in a specific setting to »see what happens« or to develop a narrative that is realistic and can plausibly be taken to be exemplary for »real-life« cases – he rather declares that his intent was to radicalize the story to an extreme. In fact, it would be difficult to imagine that the behavior of these three physicists, who fake their insanity in an idyllic Swiss sanatorium and become murderers, could exemplify what is actually going on among scientists who struggle with the potential risks of the theories they produce.

---

\(^9\) The English translation of points 2 and 3 (which in German read »2 Geht man von einer Geschichte aus, muß sie zu Ende gedacht werden./ 3 Eine Geschichte ist dann zu Ende gedacht, wenn sie ihre schlimmstmöglich Wendung genommen hat.«) has »conclusion« for »Ende«. This translation is definitely correct. It seems important to note, however, that unlike the English term »conclusion«, the German word »Ende« does not entail the idea that we are dealing with the final step in a line of reasoning, the conclusion of an argument.
That this plausibility is not a concern for Dürrenmatt can be told already from the setup of the initial scene: an inspector comes to a crime scene, but instead of developing interest in the case and making first steps towards a solution, he gives up – for he knows already who the murderer was and that he cannot be taken responsible for his action – and he asks in an indifferent tone whether he could smoke and have some brandy. It is also reinforced by above-quoted points 4 and 5, where the author explicitly appeals to accident and effect. The bizarre initial setting of the play and the surprising twists the plot takes aim at attracting the spectators’ attention and not at developing a scenario that can be considered exemplary for relevant, real-life scenarios. Therefore, Dürrenmatt’s play is very different from thought experiments we know from science or philosophy. Scientific thought experiments do not depart from highly bizarre initial conditions; philosophical thought experiments sometimes do (think of Putnam’s famous Twin Earth example), but even here, the narrative needs to be plausible and accident has to be reduced to a minimum for the thought experiment to be credible. Thus, if Dürrenmatt’s play was a thought experiment, we would have to conclude that it was conducted badly.

The play is cognitively relevant, though, and it can add to our understanding. In fact, it does so on the basis of the features that disqualify it as a thought experiment: The unusual initial setting and the surprising twists make sure that the spectators get drawn into the play more and more and so pushes them to ponder about questions they might not have considered so far. In other words, the unrealistic setting and the implausible course of the plot point the spectator to very real problems of our contemporary societies. Dürrenmatt hints at this in his last point:

21. Drama can dupe the spectator into exposing himself to reality, but cannot compel him to withstand it or even to master it. (Dürrenmatt 1973, appendix)

The second part of point 21 reveals another aspect that seems relevant to me: Dürrenmatt’s play does not really argue for a clear-cut hypothesis. He does not offer a simple or straightforward solution, but he rather presents the spectator with a paradoxical situation – and he does so in the most dramatic of all possible ways for he knows that this way his play will be most effective. If my reading of the play is plausible, the cognitive value of the work does not lie in its unfolding of a thought experiment (and not in arguing for a conclusion) but rather in its power to guide the spectators’ attention towards most pressing features of our life-world and so »dupe« them to reflect about issues they might not have taken into consideration so far. Dürrenmatt knows well enough that he cannot force his own views on the spectators – overly didactic plays often achieve the inverse result – but he can bring them to come to their own conclusions on the matters that he considered relevant and that might have gone without notice otherwise.
Conclusion

The aim of this paper was to show that the notion of thought experiment can be helpful for our understanding of the cognitive value of some works of literary fiction but also to insist that this analogy has its limits.\textsuperscript{10} We should appreciate the fact that works of literature have other means to broaden the readers’ horizons and to offer them cognitive rewards. My aim was to suggest that this is true even for works that might actually be read as thought experiments, like Dürrenmatt’s *The Physicists*.

Against the background of the current debate in the philosophy of literature it seems important to remind ourselves that our considerations on the cognitive value of literary works of art should always focus narrowly on the distinctively literary features of these works. Literary works of art can guide the readers’ attention, and this guidance is very often – though, of course, not in all cases – the mechanism on the basis of which the work can be cognitively valuable to the reader. In most cases it does so not by transmitting true propositions onto the reader but by inviting her to imagine and reflect, to consider certain topics and come to her own conclusions. And very often they do so on the basis of the work’s aesthetic features. Most importantly, we must never forget that readers are autonomous individuals who cannot be prescribed what to believe. Authors who want to change the cognitive perspective of their audience can at best invite it to reflect on particular topics and propose a certain perspective on things. Whether or not readers are willing to share the proposed conclusions or refuse them and come to their own conclusions is up to them.

References


\textsuperscript{10} An interesting example of a literary text that does work as a thought experiment is Jorge Luis Borges’ short story “Pierre Menard, Author of the Quixote.” I discuss this point in more detail in Huemer (2019).


