Internal and External Validity in Thought Experiments

Proceedings of the Aristotelian Society (2016)

James Wilson

Abstract. This paper develops an account of rigour in the use of thought experiments in ethics. I argue that there are two separate challenges to be faced. The first is internal validity: is the thought experiment designed in a way that allows its readers to make judgements that are confident and free of bias about the hypothesis or point of principle that it aims to test? The second is external validity: to what extent do ethical judgements that are correct of the world of the thought experiment generalise to a wide variety of other contexts, including ethical decisionmaking in the actual world? Ensuring external validity is the harder and more important problem of rigour, yet it is one that few philosophers have even noticed, let alone begun to solve.

I

Introduction. It is no secret that making wise ethical choices is difficult. It might reasonably be thought that one of the tasks of moral philosophy is to reduce this difficulty by means such as distinguishing and categorising cases, and devising or discovering moral principles that can guide moral judgements. One of the key ways that distinguishing, categorising and generation of principles has been performed within recent ethics literature in analytic philosophy is through the use of thought experiments. Our task in this article is to advance an initial answer to a question that needs to be answered convincingly if such an approach to ethics is to be vindicated: what is it for the use of thought experiments in normative ethics to be rigorous?[1]

Thought experiments as I understand them here are toy ethical cases that are designed to simplify an ethical problem along a number of dimensions, thus

1 Elements of my analysis bear on the use of thought experiments outside of normative ethics, but I shall have little to say about their use in other areas of philosophy. I confine myself to normative ethics because such thought experiments have an interesting and problematic feature that occurs less frequently elsewhere. Whereas thought experiments in other areas of philosophy are predominantly used to establish modal claims or conceptual entailments, thought experiments in normative ethics aim to establish what ought to be done in a sparsely described possible world and use this as a way of giving insights into what should be done in the actual world. Our key question is whether, and if so how, they can provide such insights.

making the problem more philosophically tractable.[2] I have chosen a classic thought experiment from Rachels (1975) as an initial example. Rachels aims to answer the question whether killing is, in itself, worse than letting die. He draws the reader's attention to the fact that he has constructed a pair of cases that "are exactly alike except that one involves killing whereas the other involves letting someone die":

> In the first, Smith stands to gain a large inheritance if anything should happen to his six-year-old cousin. One evening while the child is taking his bath, Smith sneaks into the bathroom and drowns the child, and then arranges things so that it will look like an accident.
>
> In the second, Jones also stands to gain if anything should happen to his six-year-old cousin. Like Smith, Jones sneaks in planning to drown the child in his bath. However, just as he enters the bathroom Jones sees the child slip and hit his head, and fall face down in the water. Jones is delighted; he stands by, ready to push the child's head back under if it is necessary, but it is not necessary. With only a little thrashing about, the child drowns all by himself, "accidentally," as Jones watches and does nothing. (Rachels, 1975, p.79)

Rachels reasons that examining two cases that differ in precisely one respect will allow the reader to isolate the ethical difference (if any) that there is between killing and letting die. His language and approach suggests that the use of the thought experiment should be understood in the same light as a scientific controlled experiment—and this is a theme that I shall take forward in what follows.

Rachels's thought experiment belongs to a class of thought experiment that I shall call an *interventional sequence.* In an interventional sequence, a base case is presented, and then modified in one or more further cases with the aim of discovering the effect that the altered feature has on ethical judgements. We can distinguish interventional sequences from *one-off* thought experiments, where a single scenario is presented for ethical consideration without attempting to modify that scenario in a controlled way.

Some thought experiments are advanced as clear or paradigm cases for the appropriateness or inappropriateness of a particular ethical judgement. Peter Singer's Shallow Pond example provides a memorable example: "if I am walking past a shallow pond and see a child drowning in it, I ought to wade in and pull the child out. This will mean getting my clothes muddy, but this is insignificant, while the death of the child would presumably be a very bad thing." (Singer, 1972, p.231) Call these *clear cases.* Clear cases can be either one-offs, or form part of interventional sequences.[3]

---

2 To anticipate some of our conclusions, thought experiments fall on a continuum from the more austere to the more richly described. All thought experiments involve selectively presenting some, rather than other, elements of a world for contemplation. This selective presentation of a world is also a feature of literature, opening the possibility that works of literature can be thought experiments. From now on, when I talk about thought experiments, this should be understood to be confined to thought experiments in normative ethics.

3 Rachels's interventional sequence relies on the assumption that both Smith and Jones present clear cases of morally wrong behaviour.

Other thought experiments are presented as *problem cases,* in which what should be done is thought to be either by its nature difficult or unclear; or where the case creates a problem for consistency with other judgements that are commonly held. Problem cases can also be either one-offs or form part of interventional sequences. Bernard Williams's famous thought experiments critical of utilitarianism were one-off problem cases. For example, George the Chemist, who is offered a way out of long-term unemployment through a job designing chemical and biological weapons, is designed to raise a question about whether a moral theory that gives no intrinsic weight to integrity can adequately account for some fundamental features of moral life. (Williams 1973, pp.97–100) The trolley literature, following Foot (1967) and Thomson (1976), presents a complex interventional sequence of problem cases about the conditions under which it is permissible to cause harm to one person in the course of averting the death of a greater number.[4]

II

Internal and external validity. What kinds of support can thought experiments in normative ethics provide for normative claims? What is it for them to be rigorous? The view that the terminology of thought experiment suggests, and which is supported by many practitioners of thought experiments in normative ethics, is that a thought experiment is a kind of experiment. Just as scientific experiments are rigorously controlled in order to answer precise research questions, while minimising the risk of contamination of the results by confounding factors, so should thought experiments be.[5]

If thought experiments are a kind of experiment, and depend for the epistemic force of their results on the rigour of their research design, it is worth beginning any search for rigour in thought experiments with the very much better established literature on experimental research design.[6] Within this

---

4 For the sake of completeness, there can also be mixed interventional sequences —either where an initially clear case is modified progressively in order to shed light on what would otherwise be thought to be a problem case (for example, by piecemeal adding and analysing additional factors to a clear case), or where an initially problem case is modified progressively in order to show how it can be thought of as a clear case. Unger (1996) provides an example of such a mixed approach.

5 Kamm puts it thus: "Real-life cases often do not contain the relevant—or solely the relevant—characteristics to help in our search for principles. If our aim is to discover the relative weight of, say, two factors, we should consider cases that involve only these two factors, perhaps artificially, rather than distract ourselves with other factors and options." (Kamm, 1993, p.7)

6 Someone might wish to deny that thought experiments are a genuine kind of experiment, or might claim that whilst thought experiments are a kind of experiment, rigour in thought experiments is *sui generis*, and thus there is

experimental research design literature, it is common to make a distinction between *internal* and *external* validity. Internal validity is a measure of the quality of the research design: an experiment is internally valid to the extent that it is designed in such a way that it correctly measures the causal effect of an independent variable or variables on one or more dependent variables.

The design of experimental trials is the subject of a massive literature. For our purposes, it will be sufficient to point out a few major factors from the literature on randomised clinical trials (RCTs). In an RCT, eligible patients are allocated at random into either an intervention group or a control group. In testing a new intervention against a control, and allocating patients at random to one of the experimental arms, many potential confounding factors are ruled out. Other potential confounding factors such as the placebo effect are ruled out by single or double blinding the trial, so that those giving or receiving the intervention are unaware of whether it is the intervention or the control that is applied. Quite apart from this, there is an elaborate set of requirements for determining sample size (to ensure that, given the expected effect size, the population size is large enough to detect it within set confidence intervals); and ensuring that the endpoints used for analysis are the same as those declared in the trial protocol (to reduce the risk that accidental correlations are misdescribed as causal). Overall, a good summary would be to say that a trial has internal validity (and would meet the methodological threshold standard for publishability) only to the extent that it has been carefully and systematically designed with sufficient care to give a high degree of confidence that the results reported accurately measure the nature and the effect size of the intervention tested.

Even if an clinical experiment is internally valid, it may tell us little about whether the same intervention will work in other circumstances. This is because an RCT, if well designed, establishes a conclusion with a high degree of confidence *about a particular population*: it is in the nature of an RCT that the rigour in experimental design that is necessary for internal validity sets limits on how, if at all, the results can be extrapolated to other contexts. (Cartwright, 2007) Thus, if the population on whom the intervention was tested are not representative of those to whom we now wish to apply the intervention, then the implications of the research for the target population are unclear. To give a

---

nothing to be learned from experimental research design. Claiming that thought experiments are not supposed to be rigorous in the way that scientific experiments are does not by itself provide support to the claim that philosophical thought experiments *are* rigorous. Moreover, the problems of internal and external validity that will be uncovered in the next sections are not dependant on the claim that thought experiments are scientific experiments, but derive from the fact that thought experiments in normative ethics provide a simplified model of a kind of scenario that might happen, and use this model as a way of reflecting on and gaining understanding of the actual world. The challenge of rigorously moving between the world of the thought experiment, and the actual world is inextricable from the use of thought experiments in ethics, and cannot be dispelled by a disavowal of scientific methodologies.

common example, if a drug was tested only on otherwise healthy young men between the ages of 21-30, it will be far from clear whether the same results should be expected if the drug is given to a frail elderly woman with multiple co-morbidities and who is already taking many other medications. (Rothwell, 2005)

To sum it up in a slogan, as Cartwright (2013) does, an RCT can show that an intervention worked *somewhere* but not that it will work *here.* A trial is said to have *external validity* if its results are applicable to a wide variety of other contexts, and in particular if there is reason to think that its intervention will work *here.*[7]

Making the distinction between internal and external validity allows us to see that there are *two* crucial questions about rigour in thought experiments. The first, on which Section III focuses, is the nature of internal validity in ethical thought experiments. As a first approximation, we might say that, by analogy to a clinical experiment, a thought experiment is internally valid to the extent that it allows its readers to make judgements that are confident and free of bias or other confounding factors about the hypothesis or point of principle that it aims to test. However, as we shall see (Section IV), the very idea of internal validity in thought experiments is significantly complicated by the fact that thought experiments are a type of fiction, and that making judgements about cases presented in thought experiments is to respond to fictions.

The second crucial question, which is the subject of Section V, is external validity. To what extent do ethical judgements that are correct of the world of the thought experiment generalise to a wide variety of other contexts, including ethical decisionmaking in the actual world? To anticipate the results of this analysis: just as in the case of clinical trials, designing for internal validity in thought experiments is vital. But just as in clinical trials, there is no easy route from internal to external validity. The paper concludes by suggesting two potentially complementary ways forward in addressing (though not solving) the problem of external validity.

III

Internal validity. I distinguished between four types of thought experiments: (1) one-off clear cases, (2) one-off problem cases, (3) interventional sequences of clear cases, and (4) interventional sequences of problem bases. One-off clear cases are standardly introduced for rhetorical purposes as the uncontroversial starting-point from which more contentious results will be derived. There is usually little attempt to establish that the judgements about the clear case, which

7 Internal validity is a necessary condition for external validity. The distinction between internal and external validity could thus be thought of as analogous to the distinction between validity and soundness in arguments. An experiment that leads to results that happen to be widely applicable, but does so despite a lack of rigour in its research design would be equivalent to an invalid argument with a true conclusion.

it is assumed that the audience also shares, are in fact correct. Of course, there is ample room for scepticism even about the use of one-off clear cases — the author might be mistaken that something is a clear case, or about what it is a clear case of, for instance. But given the rhetorical role that the one-off clear case plays, little of philosophical importance turns on the precise way in which it is presented. To the extent that there is a clear case to be captured, and an author's thought experiment currently fails to do this, the author's thought experiment can easily be swapped out or fixed up. As the focus of this article is on thought experiments as a way of *improving* moral insights, I shall have little more to say about one-off clear cases.

Problem cases require much more precision in their setup. By their nature, they are likely to be more controversial and more surprising than clear cases, and less likely to lend themselves to purely rhetorical use. The case setup needs to do at least the following three things to have a plausible claim to internal validity. First, what in the context of a legal analysis would be described as the *facts of the case* need to be described clearly, economically and consistently. It should be clear, for example, whether the problem is supposed to arise from the perspective of an omniscient narrator, or from the perspective of a fallible and limited protagonist. While adding colour may make the case enjoyable to read, this should not come at the expense of confusing or needlessly complicating the scenario presented. Second, the case needs to do more than to point out the conceptual possibility of a problem; it needs to be able to allow the reader to appreciate the problem as a concrete one occurring in the context of a possible scenario. Third, the problem and analysis needs to be genuine. Insofar as the problem relies on the set of choices available to the actors within the thought experiment being significantly reduced, this needs to be specified and at least justified (where justification might include mere stipulation).

Internal validity in interventional sequences of clear cases is structurally similar to internal validity in controlled clinical trials. In both cases, there needs to be a control case and one or more interventional cases, and the interventional cases must differ from the control case only in the ways described in the methodology. Thus in both cases, researchers need to be able to specify what research question is being tested; and to specify which variables are being modified between the cases, and which are being kept fixed. Thus, for example, if the thought experiment requires comparing two cases which differ only in a single respect, readers need to be confident that the cases *do* only differ in this one respect. In addition, the author will need to control for potential ordering effects in the presentation of cases.[8]

Attaining internal validity in interventional sequences of problem cases will require their authors to combine the lessons of internal validity of both one-off problem cases and interventional sequences. A genuine problem needs to be described clearly, economically and consistently in enough detail for the reader to be able to see the problem as arising in a particular context. This base case then needs to be modified in a controlled way. It is vital that what is kept the

8 See for example, Schwitzgebel & Cushman (2015).

same and what changes between the cases is explored rigorously both at the level of description of the cases, and at the level of moral analysis. This will frequently be challenging: what makes something a problem case is either an intrapersonal clash between intuition about the case, and the deliverances of theory; or an interpersonal clash between different individuals' judgement about the case. Given this level of uncertainty, there may well be disagreements about what is morally relevant in the case as described and so about which features of the case should be kept constant, and which modified in order to make progress.

A surprisingly large number of thought experiments in normative ethics fail to reach these basic requirements of internal validity. For example, here is Nozick introducing the problem of innocent threats:

> If someone picks up a third party and throws him at you down the bottom of a deep well, the third party is innocent and a threat; had he chosen to launch himself at you in that trajectory he would be an aggressor. Even though the falling person would survive his fall onto you, may you use your ray gun to disintegrate the falling body before it crushes and kills you? (Nozick, 1975, p.34)

Nozick's thought experiment is memorable but problematically unclear. The presentation of the case conflates the perspective of the omniscient narrator with that of the person at the bottom of the deep well trying to work out how to respond to a fast-falling human body. It is possible to view such a scenario from the perspective of an omniscient narrator who is able to state with certainty all the facts relating to the case, or from the perspective of the time and information starved person dealing with the situation, but not both simultaneously.[9] Quite apart from this, Nozick's attempts to enliven the case make it needlessly difficult to interpret. Why is the person at the bottom of the deep well (have they themselves been thrown down, or are they there voluntarily?) If the well is that deep, would the person at the bottom be in a good position to judge in good time that what has been dropped in is a human body? Why introduce the idea of a ray gun, which suggests a science fiction scenario, without further contextualising this?

In other cases, something that the author thinks is imaginable may not be concretely imaginable, or be imagined very differently by different readers. For example, Kamm (2006, p.352) gives the Reach Case, providing only the following description "Suppose that I stand in a part of India, but I have very long arms that reach all the way to the other end of India, allowing me to reach a child who is drowning in a pond at a great distance". Kamm intends that the Reach case will allow the reader to see that it should be treated like a case where the child is near. However it is far from clear how one is supposed to imagine the case. It seems from the context that Kamm intends that the extra-long arms be an integrated part of the individual's body. This raises a number of challenges for

---

9 Given the described physical constraints, it seems unlikely that a person who actually was at the bottom of the deep well would be in a good position to judge (a) if they would definitely be killed by the falling body, or (b) whether the person falling would definitely survive the fall if not disintegrated by the ray gun.

coherently imagining the case: is the rest of the person supposed to be scaled up in proportion to their thousand-mile long arms? The context suggests not. But if we are supposed to imagine an ordinary-sized human being with arms a thousand miles long, how is their anatomy supposed to work? How much would thousand-mile long arms weigh? How would the weight be supported?

As Elster (2011) discusses, readers need to be able to imagine in sufficient detail not only any outlandish elements of the thought experiment such as thousand-mile long arms, but also the implications of these outlandish elements for ethical norms and practice within the world of the thought experiment. Merely stipulating that the world of the thought experiment is different in various ways from the actual world, but not reconstructing the rest of the world of the thought experiment with this, is to fail to take seriously the case seriously as a genuine case.

Some audiences have accused this line of objection to the setup of thought experiments of being uncharitable: in context what Nozick or Kamm intends is clear enough. However, the appeal to charity in interpretation sits a little uncomfortably with the idea that the rigour of a thought experiment comes through its methodology, as we can see by comparing with the case for RCTs.

IV

Reproducibility, fiction and thought experiments. Experimental research design places the idea of reproducibility at its core. It is a basic requirement of science that, in order for an experiment to be publishable, its methodology must be written up in a way that allows the reader to appraise its internal validity, and to allow a suitably skilled team to reproduce the results. Readers will not usually try to reproduce the results obtained (given constraints of time, resources, and equipment), and only a small percentage of experiments will ever be reproduced, but it is core to the purpose of reporting methods that the results would be such as to be reproducible given the statement of the methods.[10]

If a thought experiment is supposed to be reproducible by the reader, it is vital that the author presents the reader with everything she would need to run the thought experiment herself. Where the author's scenario is indeterminate or under-described, and there are a variety of ways of filling in or imagining the stated scenario, and these different ways could plausibly have different ethical implications, then the author has not done enough to ensure all readers are attending to the same case. In that instance, the case is not reproducible: the scientific equivalent would be leaving some core elements of the methodology indeterminate, so that it was unclear for example, what dose of a particular drug was given or what the inclusion and exclusion criteria for the trial were.

Reproducibility in science requires not just that an experiment with the same essential features can be set up on the basis of the described method, but also

10 For some challenges for reproducibility, see for example Stodden (2014).

that the results can be replicated.[11] Thought experiments, as they are used in normative ethics, are intended to give rise to judgements about reasons, rather merely mechanistically caused responses. So while it is clear that thought experiments require that the case is reproducible, in the sense that it is clearly and consistently described, the degree to which reproducibility of judgements is required is less clear.

Case reproducibility in thought experiments faces a deep, but obvious, problem: thought experiments are a kind of short fiction.[12] Just as in longer fictions, the author presents a world in which certain events occur and choices are made, and as readers we are asked to attend to this world.[13] Just as in longer fictions, only a very small number of the questions that could be asked about this world are answered by the materials with which the author has provided us: the rest are left indeterminate. Knowing this unavoidable ambiguity, writers of literary fiction often choose to thematise it: rather than attempting the impossible task of presenting a world in such detail that all questions a reader might have about it will be answerable, they present a world that is deliberately spare, fragmentary or filtered through multiple incompatible perspectives. This way, each reader completes the world of the text on the basis of what he or she brings to it.

Writers of thought experiments also unavoidably depict worlds that are under-described. But where allowing each reader to complete the world in his or her imagination may be a virtue from the perspective of the literary writer, it is far from clear that it is a virtue if in constructing the thought experiment its author is attempting to conduct, or get the reader to conduct, a controlled experiment. If the methodology of thought experiments consists in getting individuals to interpret and to make judgements about very short fictions that are indeterminate in a large number of aspects, how is reproducibility to be ensured?

This problem of underdetermination is potentially so intractable for the use of thought experiments in ethics that it usually ignored rather than squarely addressed. I take it that the problem is usually supposed to be ameliorated by a

---

11 The discovery of widespread failures of reproducibility of results has led to a perceived crisis in disciplines such as psychology. (Pashler & Wagenmakers, 2012)

12 As Elgin explains, “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 Jorge Luis Borges. If the minimalist stories of Borges are genuine fictions, there seems no reason to deny that thought experiments are too.” (2014, p.230)

13 See for example, Kamm, “This may be just an autobiographical fact, but I don’t really have a considered judgement about a case until I have a visual experience of it. I have to deeply imagine myself in a certain situation, with an open mind… What I am saying is that, in order to have a judgement about a case, you really have to situate yourself in the case.” (Interview, in Voorhoeve, 2009, p.22)

convention of *authoritative authorial ethical framing*. On this convention, the case raises the ethical question or questions that the author of the thought experiment says it does; the ethical issue that the case raises is not subject to dispute. To further spell out the implied convention, the author of the thought experiment has, by definition, specified all the elements of the case that are morally relevant. No morally relevant differences other than those that have been stipulated by the framer of the thought experiment apply to the situation. Although each reader will fill out the details of the case in their own way in imagining it, each reader may only add colour and detail that is morally irrelevant.[14]

Thought experiment designers further attempt to finesse the problem through the use of an omniscient authorial voice, able to take in at a glance and to relate events in their essentials. The voice is able to tell us clearly and concisely what each of the actors within the thought experiment is able to do, their psychological states and intentions. The authorial voice will often stipulate that choices must be made from a short predefined menu, with no ability to alter the terms of the problem. For example, the reader may be presented with two choices: to pull a lever, or not to pull it. The world of the thought experiment may also be stipulated to operate according to laws that are plainly false of the real world. In this case, responses that would be likely to be effective in real world analogues of the thought experiment may be stipulated not to work, and other responses, which would be unlikely to be effective in real world analogues, may be stipulated to be effective.

Such techniques face a tension: insofar as the case is very sparsely described and is stipulated to have only the facts or features that its author says, then there is little to imagine and it is difficult for its readers to experience the case as a genuine case. But the more difficult it is to experience the case as a genuine case, the less it is plausible to claim that thinking about the case has added something additional to thinking about the problem in purely abstract terms, and the less confident we should be that the judgements that arise from the case have clear implications about the actual world. Alternatively, the richer and more realistic the case, and the more space the writer gives the reader as interpreter to decide what is salient, the easier it is for the reader to experience the case as genuine; but at the same time the less plausible it is to think of the case as akin to a controlled scientific intervention, and the more scope there is for disagreements in judgements about the case.

Failures of reproducibility of judgements are fatal for purported clear cases. A case fails to be a clear one if, given a perspicuous description of the case, there is a significant degree of disagreement on what should be said about it. Persistent

14 Applying this convention to Rachels's cases of Smith and Jones, one would be free to imagine Smith being of a variety of different ages, and as having a variety of reasons for wanting the inheritance; but any features attributed to Smith must also be attributed to Jones, and the reader is not free to imagine that the two have a feature that would call into question their moral responsibility for the killing or letting die (neither is a psychopath, or is subject to mind control).

disagreements about what should be done do not fatally undermine the usefulness of a problem case, but they do raise deep questions about the work that such thought experiments are supposed to do in ethical reasoning. There are at least three possibilities. First, failure of reproducibility of judgements could show that the thought experiment is flawed in its design and description. Second, there may be persistent blameless disagreement even among competent moral agents who have the same understanding of the ethically relevant features of the case — caused for instance by the factors that Rawls isolates as the "burdens of judgment". (1993, pp.54–57) Third, some might have greater expertise than others in responding to thought experiments. On such a view, if nonexperts fail to reproduce expert judgements about thought experiments, this is no more of a problem than if nonexperts cannot reproduce experimental results that require advanced lab skills and years of specialist training.[15]

Regardless of the conventions that are in play for interpretation of thought experiments, ethical judgements elicited by a thought experiment apply in the first instance to the world of the thought experiment. Just as with other fictions, it is an open question how ethical judgements that can be correctly made about a fictional world bear on the actual world. Obviously, if ethical judgements that were correct of the world of the thought experiment applied only to the world of the thought experiment and had no implications for the actual world, analysis of thought experiments would not be able to contribute to resolving the difficult ethical problems that are the point of departure for ethics in the first place. So I take it that a defence of thought experiments as a methodology in ethical research needs to be able to show not just that we are able to make helpful and accurate ethical judgements about thinly described fictional cases, but that such judgements bear in an enlightening way on ethical judgements about the real world. However, to say the least, it is unclear why it should be the case that we can make wise decisions about complex real life cases by rigorously analysing cases that are simpler (often radically simpler) than the real life cases that we are aiming to wisely resolve. (Dancy, 1985, p.166)

15 In response to the worry that some (perhaps many), say that they cannot form a confident judgement or even form a judgement at all about cases that are baroque or distant from everyday experience, Kamm suggests that, even among philosophers, there may be relatively few who have the wherewithal to be able to reproduce her judgements about particularly complex cases: 'Having responses to complex and unfamiliar cases requires that one see a whole complex landscape at once, rather than piecemeal. This often requires deep concentration. Only a few people may be able to respond to a complex case with a firm response… the "Princess and the Pea" is the fairy tale best associated with the method I describe: it tells of someone, despite much interference, who cannot ignore a slight difference in a case that others may never sense.' (Kamm, 1996, p.11)

V

The Problem of External Validity. Thought experiments can lack external validity in at least two ways. First, if the ethical judgments that can be established as appropriate in the world of the thought experiment depend on features of the normative context that are not shared in other normative contexts. Call this *normative contextual variance*. Second, if the ethical judgements that can be established as appropriate in the world of the thought experiment presuppose causal structures that are relevantly different from those that are present in other contexts. Call this *non-transferability of causal structures.*

Normative contextual variance. Rachels's case contrast strategy presupposes that if there is no moral difference between what Smith does and what Jones does in the precisely equalised cases, then there is no intrinsic moral difference between killing and letting die. More broadly, the underlying thought seems to be that if there has been a proper set up and analysis of the two cases (i.e. if there is internal validity in the design of the thought experiment), then the fact that a given moral feature makes a difference in the thought experiment shows that it makes a difference everywhere. Kagan (1988, p.12) describes this as the *ubiquity thesis.* If the ubiquity thesis were true, then external validity would come for free with internal validity.

However, it is at best unclear whether the ubiquity thesis should be accepted. It is important to notice how strong the thesis is. First, the thesis requires that if we can find *any* pair of precisely equalized cases, and show that the feature which is different between the two cases either matters, or does not matter, then this result applies to all other contexts. But ethical principles and considerations are often claimed to interact with one another in holistic ways. On this view, there are scenarios in which ethical considerations that favour acting in certain ways in many or most cases no longer provide a reason in favour of acting in that way, and may even change polarity and provide a reason against acting in that way. For example, in usual circumstances, the fact that doing X will provide someone else with pleasure speaks in favour of it, but there are readily imaginable scenarios where the fact that something will create pleasure for someone would count against it (suppose the pleasure is sadistic).

Many moral philosophers, and many nonphilosophers thus endorse what Kamm describes as the Principle of Contextual Interaction, namely that a moral property can "behave differently in one context than in another". If the Principle of Contextual Interaction is correct, then it might be the case that "in some equalized contexts, a harming and a not-aiding will be judged as being morally equivalent, yet in other equalized contexts, they will not be". (Kamm, 2006, p.17)

The ubiquity thesis licenses inferences of a universalistic kind that, from the perspective of the Principle of Contextual Interaction, look to be obviously unsafe. For example, the ubiquity thesis licenses the inference that if we can find one pair of cases (such as Smith and Jones) which are precisely equalized and differ only one feature, and that there is no moral difference between these cases,

then *for any pair of cases* that are precisely equalised and differ only in this respect, there will be no moral difference between the cases.[16] The fact that the ubiquity assumption is itself controversial and has been subjected to many purported counterexamples, suggests that it would make work in normative ethics less rather than more rigorous if writers in normative ethics were to presuppose its truth in using thought experiments. Failures of external validity as a result of normative contextual interaction must be expected and planned for if the use of thought experiments in ethics is to be responsible.[17]

Non-transferability of causal structures. In the case of scientific experiments, failures of external validity arise from a nontransferability of causal structures: something that was presupposed in the background causal structures for the experiment no longer holds in the environment to which the intervention is transferred. This is a particular problem for interventions at a policy level, where there are often indefinitely large number of background features presupposed by successful randomised controlled interventions.

Cartwright and Hardie give an example of the California class-size reduction program. Following a well-conducted RCT in Tennessee, which showed that reducing class sizes was effective in improving reading scores, an attempt was made to try the same intervention in California. However, in California, the intervention failed to improve reading scores. The program is thought to have failed because some background features of the Tennessee intervention were not replicated in California:

> In Tennessee, the project involved only schools that had available space. In California, there was often not enough spare space. So sometimes space was found, but it was not as good as existing classrooms. And it was taken away from other activities that might be thought important for student achievement—special needs, music and arts, athletics, and child care programs. Second, Tennessee had no shortage of qualified teachers to staff the reduced size classes. But California had to hire an additional 12,000 teachers. And many of these were unqualified. (Cartwright & Hardie, 2012, p.66)

If we take thought experiments in any way seriously as a basis for possible action, we will have to contend with exactly the same problem. This problem arises particularly acutely in cases where the thought experiment purports to model a real-world choice context, but has a significantly different experiential, psychological, causal or epistemic structure from the real world context.

---

16 This leaves such attempts to establish the equivalence of harming and not-aiding under significant threat of counterexample: "if we can find even one set of comparable cases in which a harming is morally worse than a not-aiding, we rebut the Equivalence Thesis, for while a single positive instance cannot prove a universal claim, a single negative instance can defeat it." (Kamm, 2006, p.17)

17 It would be precipitate for philosophers to give up on interventional sequences of cases on this basis. Sorensen argues that, so long as these interaction effects are rare, all that is required is a policy of "mild vigilance", and a commitment to follow-up with further experiments where interaction effects are suspected. (1992, pp.172–3)

(Bauman et al., 2014; Wilson, 2009) Thought experiments about risk are particularly problematic in regard. In the world of the thought experiment, it will often be stipulated that there is certainty of effect: each of the defined choices will bring about its stipulated effect with certainty. It will be possible to identify in advance who will benefit, and who will lose, from each of the predefined choices. Even where the choice is stipulated to bring about the desired effect a certain percentage of the time, these stipulated probabilities are entirely accurate. Barely any real world cases of risk imposition have this structure.[18]

Here is an example from a recent article by Frick, which in fact begins by acknowledging that contractualist accounts of harm have so far problematically focused on "*certain* harms to *known* individuals" (2015, p.178), where such cases are in fact rare. Frick aims to extend contractualist analysis to the types of risk that occur in social policy, centring his analysis around the following case:

> Mass Vaccination (Known Victims): One million young children are threatened by a terrible virus, which is certain to kill all of them if we do nothing. We must choose between mass producing one of two vaccines (capacity constraints prevent us from producing both):
>
> Vaccine 1 is certain to save every child's life. However, the vaccine will not provide complete protection against the virus. If a child receives Vaccine 1, the virus is certain to paralyze one of the child's legs, so that he or she will walk on crutches for the rest of his or her life.…
>
> Vaccine 3 is sure to allow 999,000 children to survive the virus completely unharmed. However, because of a known particularity in their genotype, Vaccine 3 is certain to be completely ineffective for 1,000 identified children. These doomed children are sure to be killed by the virus if we choose Vaccine 3. (Frick, 2015, pp.181–3)

The case described differs in various salient ways from the causal mechanisms by which infectious diseases are spread in the actual world. First, in the actual world exposure to the infectious disease pathogen is a necessary, but not a sufficient condition for developing a clinically significant infection: whether a clinically significant infection follows depends on the interaction of features of the host, environment and pathogen. Second, in any real-life scenario there would be a distribution of degrees of severity of clinical symptoms, rather than a sharp divide into two homogeneous groups. Moreover, in the real world vaccines are rarely 100% effective: they are likely to be more effective for some than for others; some may be allergic to the vaccine, or unable to attend on the day of the mass vaccination. So in any real world scenario, whatever is done some children will not be vaccinated. Thus, in any case that has the causal structure of an infectious disease outbreak in the *actual* world, the number of persons exposed to the pathogen, and the number of persons who will develop clinically relevant

18 As Fried (2012a; 2012b) argues, there is an additional worry about external validity and contextual variance here. Individuals' judgements about cases involving *certain* harm are often difficult to reconcile with their judgements about *risks* of harm. It is unclear why it should be assumed to be more methodologically robust to examine the ethics of risk imposition in a way that requires individuals to imagine away the everyday experiences they have of risk, than by building on these experiences.

symptoms will not be knowable in advance, but will only be able to be estimated (often within wide confidence intervals) on the basis of mathematical models.

Third, the case is stipulated to be one of a deadly virus, but no discussion is given of the mode of transmission. In any usual case of a virus that attacks human beings, and can be transmitted from one human being to another, herd immunity will be relevant. That is to say, the likelihood of being exposed to the pathogen depends on the degree to which others have been vaccinated (and so your likelihood of getting the disease if you yourself are not vaccinated is much lower in an environment where all others are vaccinated than where none are).[19] Frick's Vaccine 3 scenario elides this fact, assuming that if the vaccine is ineffective for 1,000 children then these children will be (a) sure to be exposed to the pathogen, and (b) die as a result of being exposed to the pathogen.

Fourth, it is unclear why the children for whom Vaccine 3 would be ineffective could not be isolated until the disease passes (they are stipulated to be indentifiable, and to be only 1000 out of 1 million). Overall, there are so many differences in underlying causal structure between Frick's case and any actual vaccination policy case, that even assuming internal validity for Frick's analysis of the thought experiment and leaving on one side any questions of normative contextual variance, the thought experiment has very limited relevance for actual vaccination policy decisions.

VI

Responding to failures of external validity in thought experiments. The results of our analysis place thought experiments in ethics in a significantly weaker position than randomised clinical trials with respect to internal validity, but roughly on a par when it comes to external validity. Even if a thought experiment *is* internally valid, this does not provide a strong reason to think that it is externally valid—due both to the possibility of normative contextual interaction, and to the likelihood that the actual world has a relevantly different causal structure from the world of the thought experiment.

Just as the failure of internal validity to guarantee external validity does not imply that randomised clinical trials cannot be externally valid, so the failure of internal validity to ensure external validity in thought experiments does not imply that thought experiments cannot be externally valid. So I have provided no reason to think that thought experiments should be (even could be) abandoned in normative ethics. Rather, what we have seen is that even rigorously designed thought experiments may not show what their designers think they do and that consequently philosophers need to be on the look out for, and to expect, defeaters that prevent translation from one normative context to another.

19 An exception to this general principle of vaccination would be tetanus, which is not communicable from person to person. However, given the description of the case, there does not seem to be any reason to suppose that Frick has in mind a noncommunicable virus.

The rest of the article suggests two potentially complementary ways forward in addressing (though not solving) the problem of external validity. The *translational model* is drawn from the way that the problem of external validity has been addressed within medical science. The *art model* suggests a nonscientific approach, according to which thought experiments' ability to inform ethical reflection should be conceived as akin to (or to be an example of) the way that literature can inform ethical reflection.

The translational model. How should the organisation of research in philosophy change as a result of giving adequate weight to the problem of external validity? I have argued elsewhere (Wilson, 2014) that medical research provides a useful analogue, both for seeing the problems that come with an approach that focuses only on internal validity, and also for the types of difficulties that need to be overcome to re-orient a research programme towards external validity. The following paragraphs summarise that argument.

For a period of thirty years after the second world war, government approaches to science funding, following the influential report, *Science: The Endless Frontier*, by Vannevar Bush in 1945 presupposed a linear model of scientific innovation. Such a model presupposes a sharp distinction between basic research, which is "performed without thought of practical ends…[and] results in general knowledge and an understanding of nature and its laws", (Bush, 1945 Chapter 3.3) and applied research, which makes use of basic research in order to answer practical problems. On the linear account, it was assumed that investing in basic research would (although in ways that cannot now be predicted) have significant practical payoffs in the future, though little was done to theorise how this would come about, or to optimise the processes through which this would happen. (Godin, 2006; Balconi et al., 2010)

The linear model of innovation has been thoroughly discredited. To give just two examples, the idea that advances in basic science are always necessary for improved technologies – in Bush's words, that "basic research is the pacemaker of technological progress," (1945 Chapter 3.3) has been proved false. Whilst there are obvious cases where new therapies have been developed out of a bedrock of advances in basic science (such as monoclonal antibodies), it is implausible to claim that healthcare innovation always starts from advances in basic science.[20] Second, the idea that new basic science automatically leads to changes in medical practice that benefit patients is also highly questionable, given that the history of medical science is littered with examples of failures to properly connect basic and applied science.

As a result of the shortcomings of the linear model, a consensus has grown that funders and researchers need to rethink research rigour and funding priorities in *systemic* terms: what matters is not only whether isolated individuals or labs do excellent and internally valid research, but also how the research continuum

20 Indeed, influential innovation scholars such as Kline and Rosenberg (1986, p.288) argue that innovation led by basic research is the exception, rather than the rule.

from the most basic science to the most applied fits together in a unified system. A simple top-down linear model of innovation has been replaced by a more complex translational model, which recognises that innovations move along different pathways – both from the more 'basic' to the more 'applied' but also from the more 'applied' to the more 'basic'.

My impression is that many within normative ethics, particularly those who make heavy use of thought experiments, adhere (perhaps unwittingly) to a top-down linear model. In this context, basic research would be discussion of 'pure' moral theory without any attempt to think about the applicability of moral theory to real life cases; working out what ought to be done in thought experiments would still be close to basic research, while working out should be done all things considered in real world situations would be applied research. Just as in a linear model of scientific innovation, it is assumed that insights will trickle down from more basic theory to more applied contexts, but it is not thought that insights also need to trickle up from 'applied' to more 'basic' theory.[21] A top-down linear approach in normative ethics is plausible only on the assumption that external validity does not present a serious problem. If, as we have shown, there are systematic problems in establishing external validity in normative ethics, there is no reason to think that the examination of simpler and more abstract cases will automatically lead to insights that will be helpful in responding to messier and more complex ones.

What is the right way forward? I suggest that a good start is to shift towards a translational model for normative ethics, rather than a top-down linear one. A translational model will take the search for external validity to be its core, and will rethink the use of thought experiments, from an implicit model in which achieving internal validity in highly simplified scenarios is taken to be an end in itself, to one in which highly simplified scenarios are used alongside richer and more realistic cases in the search for what should be done in the actual world. For translational ethics, as for a linear model, there will still be a continuum from thought experiments to real life cases, but the interpretation given to this continuum is transformed. From the perspective of translational ethics, we can and should range over the continuum in both directions: as we travel across the continuum in one way, more and more of the messiness and complexity of the real world is bracketed for simplicity, and in the other way, more and more of the messiness and complexity of the real world is revealed.

Translational ethics will be alive to the ways in which changing the frame of reference, or shifting a highly simplified scenario towards something more realistic can disrupt and challenge external validity. It will also be alive to the

21 See for example, GA Cohen's demand that the subject matter of political philosophy be the search for fact-free fundamental principles of justice: "facts are irrelevant in the determination of fundamental principles of justice. Facts of human nature and human society of course (1) make a difference to what justice tells us to do in specific terms; they also (2) tell us how much justice we can get; and they (3) bear on how much we should compromise with justice, but, so I believe, they make no difference to the very nature of justice itself." (2009, p.285)

ways that analysis of rich cases can themselves have important theoretical implications: for example, they may bring out problems or contradictions, or relevant values, that went unnoticed in a more abstract model. (As we saw, much contractualist writing on the ethics of risk could benefit from being complicated in this way.)[22]

The art model. Another response to the problem of external validity is to stop thinking of thought experiments according to a scientific or quasi-scientific model. On the scientific model, internal validity comes first: once internal validity is established, there is a further question about how (if at all) the judgements that are correct of the world of the thought experiment apply to the actual world. An alternative would be actively to embrace the idea that thought experiments are *fictions,* and to reconceive the way that thought experiments can provide ethical insights along the lines of the way that other fictions such a novels or dramas can. Again, I can only sketch how this suggestion could be taken forward here.

Paul Ricoeur argues that what makes art cognitively and ethically valuable is not its ability to represent the world—in the sense of providing recognisable simulacra of the world we ordinarily inhabit, but rather in its ability to "discover dimensions of experience that did not exist prior to the work." (1998, p.173) For Ricoeur, the artwork holds out a world that we can enter imaginatively, and we can use the world of the work as a vantage-point from which to look back, and restructure our understanding and experience in the actual world. It is in virtue of the world of the work's *difference* from the every day actual world that the world of the work allows us to reflect creatively on the actual world:

> …as the gap with reality grows wider, the biting power of the work on the world of our experience is reinforced. The greater the retreat, the more intense the return back upon the real, as coming from a greater distance, as if our experience were visited from visited from infinitely further away than itself. We have a sort of counterexperience of this hypothesis in the example of photography as it is practised by amateurs, when what we obtain is simply a

22 One obvious question is how this translational model differs from reflective equilibrium, which also consciously eschews a top-down approach. My answer would be that once reflective equilibrium has been purified in such a way as to avoid charges of conservatism, it is not a determinate method for moral thinking, but in Scanlon's words, "a level playing field of intuitive justification on which principles and judgements of all levels of generality must compete for our allegiance". (2003, p.151) Thus reflective equilibrium *per se* does not address the relationship between internal and external validity. There are many philosophers who take themselves to be pursuing reflective equilibrium solely by attempting to reconcile abstract theoretical claims with judgements about thought experiments. It will not surprise the reader to hear that I do not think that such an approach, when used on its own, is particularly likely to lead to external validity. There are, however, also approaches to reflective equilibrium such as Thacher (2006), which use richly described real life cases as a way of bringing particular judgements into dialogue with more abstract theoretical ones, which fit better with translational ethics as I conceive it.

> double of the real, with a return to the origin by way of only a very small loop, and, as a result, its grip on our world is infinitely less. (Ricoeur, 1998, p.176)

In one way, it might seem as if things are in good order, by Ricoeur's lights, when it comes to thought experiments: haven't we gone a long distance from the real with the positing of thousand-mile long arms or ray guns? But this would be to misunderstand Ricoeur, who talks of art as "restructuring the world of the reader in confronting him or her with the world of the work". (1998, p.173) The problem with most thought experiments is that the world of the world is so sparsely and lackadaisically described that no world that could confront the reader has been presented. The average thought experiment has a flatness analogous to that which Ricoeur ascribes to the amateur photograph, though whereas the amateur photograph reproduces the source world without transforming it, the thought experiment reproduces the theory laden assumptions of the philosopher without creating a self-standing world from which these assumptions can be properly scrutinised.

What would a thought experiment look like that would allow a fuller sense of a world from which the world of experience could be rethought? First, just as in other fictions, the world of the thought experiment needs to be fully-realised and self-consistent (or if inconsistent, deliberately so). Insofar as a state of affairs is said by the author of a thought experiment to obtain within the world, the rest of the world of the thought experiment needs to be consistent with this. By this, I mean not just factually, but sociologically consistent. To the extent that the world of the thought experiment is different from the taken-for-granted everyday world, and the author wishes the readers of the thought experiment to be able to reflect on the ethical scenario it presents, the author (like the author of a science fiction story) must think through the implications of these changes for moral agency and identity, and the broader social structures that affect it.[23]

Second, the case needs to leave enough space for the reader as interpreter to genuinely enter the world: space to decide for herself what she takes to be salient about the scenario presented and how to respond to it; and space to raise the questions that the author has either failed to ask, or has attempted to sidestep.

VII

Conclusion. It is often thought that thought experiments are a key way through which the rigour of ethical theorising is increased. However, theorists have thus far failed to adequately distinguish between two kinds of rigour in the use of thought experiments in ethics: internal and external validity. External validity is both more important and more difficult to ensure, but philosophers have tended overwhelmingly to focus on internal validity. Taking external validity seriously should lead philosophers to be far more reflexive about the limits of thinly described thought experiments. They are but one tool among many that can

23 Becker (2007, chap.7) has a discussion of the use of parables in sociology that is useful here.

contribute to the ultimate purpose of improving our ability to act wisely in the world as we find it.

References

Balconi, Margherita Brusoni, Stefano and Orsenigo, Luigi 2010: 'In defence of the linear model: An essay'. *Research Policy*, 39(1), pp. 1–13.

Bauman, Christopher, McGraw, A. Peter, Bartels, Daniel and Warren, Caleb 2014: 'Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology'. *Social and Personality Psychology Compass*. 8 (9), 536–554.

Becker, Howard 2007: *Telling about society*. Chicago: University of Chicago Press.

Bush, Vannevar 1945: *Science: The Endless Frontier*. Washington: United State Government Printing Office.

Cartwright, Nancy 2007: 'Are RCTs the Gold Standard?'. *BioSocieties*, 2(01), pp. 11–20.

—— 2013: 'Knowing what we are talking about: why evidence doesn't always travel'. *Evidence & Policy: A Journal of Research, Debate and Practice,* 9(1), pp. 97–112.

Cartwright, Nancy and Hardie, Jeremy 2012: *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York: OUP USA.

Cohen, G. A. 2009: *Rescuing justice and equality*. USA: Harvard University Press.

Dancy, Jonathan 1985: 'The role of imaginary cases in ethics'. *Pacific Philosophical Quarterly*, 66(1-2), pp. 141–153.

Elgin, Catherine 2014: 'Fiction as thought experiment'. *Perspectives on Science,* 22(2), pp. 221–241.

Elster, Jakob 2011: 'How outlandish can imaginary cases be?' *Journal of Applied Philosophy*, 28(3), pp. 241–258.

Foot, Philippa 1967: 'The problem of abortion and the doctrine of double effect'. *Oxford Review*, 5, pp. 5–15.

Frick, Johann 2015: 'Contractualism and social risk'. *Philosophy and Public Affairs*, 43(3), pp. 175–223.

Fried, Barbara H. 2012a: 'Can contractualism save us from aggregation?' *Journal of Ethics*, 16(1), pp. 39–66.

—— 2012b: 'What does matter? the case for killing the trolley problem (or letting it die)'. *Philosophical Quarterly*, 62(248), pp. 505–529.

Godin, Benôıt 2006: 'The Linear Model of Innovation: The Historical Construction of an Analytical Framework'. *Science, Technology & Human Values*, 31(6), pp. 39–667.

Kagan, Shelly 1988: 'The additive fallacy'. *Ethics*, 99(1), pp. 5–31.

Kamm, F. M. 1993: *Morality, Mortality, Volume I: Death and Whom to Save From It*. Oxford: Oxford University Press.

—— 1996: *Morality, Mortality, Volume II: Rights, Duties, and Status*. USA: Oxford University Press, USA.

—— 2006: *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford University Press, USA.

Kline, Stephen and Rosenberg, Nathan 1986: 'An Overview of Innovation'. In Ralph Landau and Nathan Rosenberg (eds.) *The Positive Sum Strategy: Harnessing Technology for Economic Growth*, pp. 275–306. USA: National Academies Press.

Nozick, Robert 1974: *Anarchy, State, and Utopia.* New York: Basic Books.

Pashler, Harold and Wagenmakers, Eric-Jan 2012: 'Editors' introduction to the special section on replicability in psychological science a crisis of confidence?' *Perspectives on Psychological Science*. 7 (6), 528–530.

Rachels, James 1975: 'Active and passive euthanasia'. *New England Journal of Medicine*, 292(2), pp. 78–80.

Rawls, John 1993: *Political Liberalism*. Columbia: Columbia University Press.

Ricoeur, Paul 1998: *Critique and Conviction: Conversations with François Azouvi and Marc de Launay*. Oxford: Polity Press.

Rothwell, Peter 2005: 'External validity of randomised controlled trials: "To whom do the results of this trial apply?"'. *The Lancet,* 365(9453), pp. 82-93.

Scanlon, T. M. 2003: 'Rawls on Justification'. In Samuel Freeman (ed.) *The Cambridge Companion to Rawls*, pp. 139–167. Cambridge: Cambridge University Press.

Schwitzgebel, Eric and Cushman, Fiery 2015: 'Philosophers' biased judgments persist despite training, expertise and reflection'. *Cognition*, 141, pp. 127–137.

Singer, Peter 1972: 'Famine, affluence, and morality'. *Philosophy and Public Affairs*, 1(3), pp. 229–243.

Smart, J.J.C. and Williams, Bernard 1973: *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

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

Stodden, Victoria 2014: 'Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency'. In Julia Lane, Victoria Stodden, Stefan Bender and Helen Nissenbaum (eds.) *Privacy, Big Data and the Public Good: Frameworks for Engagement*, pp. 112–132. New York: Cambridge University Press.

Thacher, David 2006: 'The Normative Case Study'. *American Journal of Sociology*, 111(6), pp. 1631–1676.

Thomson, Judith Jarvis 1976: 'Killing, letting die, and the trolley problem'. *The Monist*, 59(2), pp. 204–217.

Unger, Peter 1996: *Living High and Letting Die: Our Illusion of Innocence*. Oxford: Oxford University Press.

Voorhoeve, Alex 2009: *Conversations on Ethics*. Oxford: Oxford University Press.

Wilson, James 2009: 'Towards a Normative Framework for Public Health Ethics and Policy'. *Public Health Ethics*, 2(2), pp. 184–194.

—— 2014: 'Embracing complexity: theory, cases and the future of bioethics'. *Monash Bioethics Review*, 32(1-2), pp. 3–21.