The Logic of Inference of Thought Experiments in Political Theory

Adrian Blau
Senior Lecturer in Politics
Department of Political Economy
King’s College London
Adrian.Blau@kcl.ac.uk

Abstract: This paper highlights important parallels between political theory thought experiments and comparison in the natural and social sciences. Most of these similarities have not been noted before. This gives us a more precise language with which to assess the strengths and weaknesses of thought experiments. And it gives us powerful tools for improving them, by using ideas like internal and external validity, controlled comparison, omitted variable bias, interaction effects, spurious correlations, testable implications, and parsimony. Focusing on variables is the key. This helps me address longstanding debates about ‘weird’ and ‘wacky’ thought experiments. I do not wish to exaggerate the scientific parallels: there are important differences too. But the similarities raise fascinating questions about the links between political theory and political science.

Acknowledgements: Some of the ideas in this paper were presented at the Faculty of Governance and Global Affairs seminar, Leiden University, 20 May 2016; the PSA Political Methodology conference, UCL, 27 June 2016; and the Department of Political Economy research symposium, KCL, 29 June 2016. I thank participants for their comments and criticisms.

9700 words, plus references
1. Introduction

There has been much debate over whether to use thought experiments in political theory. There has been less debate over how to use them. This paper’s first contribution is that understanding how to use thought experiments casts light on whether to use them. This methodological principle – emphasizing the ‘how-to’ aspect of methods – follows the research agenda of Blau (forthcoming) in general, and Brownlee and Stemplowska (forthcoming) on thought experiments in particular.

My second contribution is to provide ammunition for both defenders and critics of thought experiments, though more for the former. This rests on my third contribution: the parallels between thought experiments and scientific comparisons. Developing Frances Kamm’s work, I relate thought experiments to controlled comparison in political science and to natural-science controlled experiments, e.g. holding all variables constant except for one, then changing that variable to see if our judgment changes.

Despite some major differences, such similarities point to my fourth contribution: suggestions to help us strengthen thought experiments. The focus on variables is crucial: successful thought experiments require control and variation/manipulation. This increases internal validity, by avoiding omitted variable bias and spurious correlation while using testable implications. It highlights external validity problems, e.g. extrapolating from parsimonious thought experiments.

The natural and social sciences thus enable my fifth contribution: a ready-made language more precise than our existing terminology. For example, Adrian Walsh discusses thought experiments in cases which are ‘very distant from our own’ (Walsh 2007, 179-80). But how ‘distant’ are experience machines or fat men drowning in rivers? The question is better posed in terms of relevance of variables, I will suggest. Similarly, Victor Tadros (2011, 7-8) talks of ‘clean’ and ‘dirty’ thought experiments. But the former fails to distinguish between parsimony and removing excess variables, while the latter seems to involve not added dirt but omitting variables, which sounds more clean than dirty. By contrast, Robert Goodin’s (1982, 8-12) powerful critique of thought experiments hits the mark through criticizing scenarios ‘too stripped down to be of any real policy guidance’ (parsimony) and obscuring ‘interactions’ between norms (omitted variable bias through interaction effects). However, his criticism of scenarios with ‘too much unrealistic or inappropriate detail’ does not distinguish between omitted variable bias (the
kidnapping in Thomson’s violinist case) and external validity (cases whose variables do not apply in real situations).

Political theorists might object to my scientific parallel. But at its most general, the scientific method – considering different interpretations, accepting the force of the better argument, using testable implications, and so on – was developed by philosophers and institutionalized by lawyers long before modern science. Scientists and statisticians have sharpened these ideas but the general approach continues. Indeed, scientific discussion shares much with deliberative democracy (Gambetta 1998, 39) and defenses of free speech (Mill, On Liberty ch. 2). I hope the ideas are not too alien.

One important set of caveats. Despite my expansive title, I do not cover all types of thought experiment, as explained shortly. I concentrate on ‘analytical’ not ‘continental’ political theory (inasmuch as this distinction is meaningful or helpful). I do not discuss scientific connections made by comparative political theorists (see Ackerly and Bajpai forthcoming) and experimental philosophers (e.g. Greene 2013). Indeed, I largely ignore experimental philosophy, partly because I distrust many thought experimenters’ stance on intuitions, as I will explain, and partly because I focus more on how to draw inferences from thought experiments than in what different people conclude. Nor do I discuss what place thought experiments should play in political theory: they are just one tool we use (Walsh 2007).

This paper is structured as follows. I offer a new framework of thought experiments (section 2), address common concerns about hypothetical cases and intuitions (section 2.1), then contrast comparative and non-comparative thought experiments (section 2.2). After explaining the logic of comparative inference in thought experiments (section 3), I note parallels and differences between thought experiments and scientific comparison (sections 4 and 4.1). Distinguishing internal and external validity (section 5) precedes a discussion of increasing internal validity by controlled comparison (section 5.1). I discuss threats to internal validity – omitted variable bias, interaction effects and spurious correlation (sections 5.2 to 5.4) – before illustrating improvements to internal validity via testable implications (section 5.5). I then tackle external validity (section 6), including the issues of counternomic thought experiments and parsimony (sections 6.1 and 6.2). The conclusion discusses the relationship between political theory and political science more generally.
2. A new framework of thought experiments

Natural scientists and philosophers have long used thought experiments (Kennard 2015). Normative theorists often use them, e.g. Plato’s Ring of Gyges, Hobbes’s state of nature, and Rawls’s original position. Less famous examples include Aristotle’s critique of Plato (is a stone hand essentially the same as a real hand?) and Mill asking if stopping a man from crossing a dangerous bridge restricts his liberty.

Thought experiments are now extremely prominent, through Nozick’s popularizing of the approach of writers like Judith Jarvis Thomson and Philippa Foot (Wolff 2013, section 27.6). For example, Thomson (1971, 48-9) defended abortion by asking if you could legitimately unplug yourself if forcibly connected to a famous violinist who would die if disconnected. Foot (2002, 23-4) started ‘trolley problems’: if a runaway train will kill five people, should we switch the track, killing just one person?

Although political theorists often imply that there is just one kind of thought experiments, philosophers distinguish several kinds. I will combine and expand two existing typologies. The first is Tamar Gendler’s (2000, 25-7) typology of factive, descriptive and valuational thought experiments, which I will call empirical, conceptual and normative. The second framework is John Norton’s (1991, 131) distinction between inductive and deductive thought experiments. But Norton only considers empirical thought experiments: I apply his distinction to conceptual and normative thought experiments too. More importantly, section 2.2 will distinguish comparative and non-comparative thought experiments. Note that I take a more subjective classificatory stance than Gendler and Norton.

Empirical-inductive thought experiments ask ‘what would happen if …?’ and answer through empirical inference. Political theory examples include Hobbes’s state of nature (what would happen without government?) and Plato’s Ring of Gyges (how would a man behave with an invisibility ring?). Our answers are guesses, extrapolating from our and others’ behavior.

Empirical-deductive thought experiments answer empirical questions by logical deduction. Galileo probably did not drop a heavy and a light ball off the Leaning Tower of Pisa to see how fast they fell but instead ran a thought experiment where deductive logic demonstrated the answer. Some famous thought experiments are presented as empirical-deductive but are arguably more, or entirely, empirical-inductive. Hobbes’s laws of nature are offered as deductively valid, but they do not all follow from his premises. Some still work if we
flesh out the argument (e.g. Martinich 2005, 95-6) but others seem like inductive inferences about how some actions might create a state of nature (e.g. Blau 2013).

Conceptual thought experiments, mostly ignored in discussions of thought experiments (though see Gendler, above, and Brendel 2004, 102-3), ask how we would *classify* certain situations. A conceptual-inductive thought experiment is Raz’s (1986, 374) hounded woman who constantly tries to escape a fierce beast on an island. She can choose freely between many options: hiding in caves, climbing trees, using camouflage, etc. But would we call this autonomy? Raz thinks not: autonomy requires a valuable set of options.

Raz’s thought experiment is what Daniel Dennett (1995) calls an ‘intuition pump’: it pumps an intuition to the surface but does not test its veracity. It helps clarify a person’s concepts but may not convince other people. This fits my partly subjective stance: thought experiments are part of a reflective equilibrium process where there may be multiple reflective equilibria. Nonetheless, conceptual-inductive thought experiments can be powerful tools in conceptual analysis (see e.g. Feinberg 1985, 10-49). (I am not aware of political theorists recommending thought experiments for probing concepts. I hope I am wrong.)

Conceptual-deductive thought experiments tell us what the conceptual answer must logically be. Consider Hobbes’s reworking of Aristotle’s example of a man who throws his goods off a sinking ship for fear that he will otherwise die. Hobbes’s contemporaries described such actions as unfree. Hobbes uses definition and logical deduction to refute this (*Leviathan* ch. 21 paragraphs 1-3, pp. 145-6). Whereas conceptual-inductive thought experiments are intuition pumps, conceptual-deductive ones are equivalent to genuine arguments. (On equivalence between thought experiments and genuine arguments, see e.g. Brownlee and Stemplowska, forthcoming section 3.1.2.) Likewise, Hillel Steiner (1975, 45-7) and G.A. Cohen (2011, 159-62) offer thought experiments about freedom that invite us to draw intuitions that are then refuted, deductively.

Normative thought experiments ask ‘is P right or wrong?,’ or ‘is P better or worse than Q?,’ or ‘should you do P or Q?’ Normative-inductive thought experiments include trolley problems and Nozick’s experience machine (discussed below): what would you do in such situations, they ask? Tversky and Kahneman’s (1981, 453) famous Asian disease scenario, like Hobbes, Steiner and Cohen, combines both normative-inductive and normative-deductive thought experiments: it asks respondents what they would do, then deductively shows the unreliability of some intuitions.
David Estlund’s (2000) voucher scenario is normative-deductive, proving that ‘political quality’ can be improved by reducing political equality. The thought experiment does not show how common this is or what the normatively right tradeoff is, but it does refute any suggestion that political equality always maximizes political quality.

Normative thought experiments can double as conceptual ones: Shelly Kagan’s thought experiments about killing versus letting die can make us consider whether letting die is killing (Kamm 2001, 72). More importantly, some people might see normative thought experiments as merely empirical-inductive TEs: we are not really saying whether P is better than Q, or whether we should pick P or Q, but whether we feel P is better than Q, whether we would pick P or Q, which could be seen as a form of ‘what would happen if …?’, i.e. ‘how would you feel if …?’. There are good reasons to resist this view (Brownlee and Stemplowska, forthcoming section 2.2), but my arguments about controlled comparison of variables apply regardless.

I do not define thought experiments: like Brown and Fehige (2016), I ‘leave the term loosely characterized, so as not to prejudice the ongoing investigation.’ Readers can agree that this paper’s examples are thought experiments without a definition.

2.1. Hypothetical versus real cases, and intuitions versus reflective appraisals

Before moving onto the scientific parallels, I will deflect two common criticisms of thought experiments and identify where the focus should be. I will first blur the apparent divide between ‘real’ examples and hypothetical/imaginary thought experiments. ‘Real’ examples never report everything: we spotlight key points and ignore others. For example, Mark Philp (1997) challenges the public/private divide assumed in corruption definitions by describing a real, borderline case of corruption. Philp ignores irrelevancies such as the individual’s age, sexuality and ideology. This stripped-down example is analogous to some stripped-down thought experiments. A real example is rhetorically stronger in Philp’s case, but the conceptual point is the same regardless. And had Philp changed the example, to pinpoint a further conceptual issue, it would be part real and part hypothetical.

So, what matters is not the hypothetical nature of thought experiments but whether their variables are sufficiently similar to real scenarios that vex us. Indeed, simple thought experiments can capture real-life variables successfully. Consider Foot’s (2002, 20-1) example of a doctor choosing between saving a woman in labor by performing a craniotomy which will kill the baby, or letting the baby be born and seeing the woman die. Alas, doctors do face such
dilemmas, and in such cases they exclude variables like the woman’s sexuality or ideology: choosing in these real scenario requires factoring out irrelevant variables, like a thought experimenter.

Thought experiments and real examples thus focus our attention on relevant variables. We can test relevance by injecting variables and seeing if it affects our judgment. Foot’s craniotomy scenario may or may not help decisions about killing versus letting die in more complex dirty hands political scenarios; Philp’s real example may or may not help us conceptualize corruption in African contexts with different public/private notions (Ekeh 1975).

Where thought experiments are equivalent to arguments (Brownlee and Stemplowska, forthcoming section 3.1.2), there is no substantive difference between an argument, a hypothetical example and a real example. Hobbes could have shown, without any example, that compulsion is consistent with liberty; he could have used a hypothetical example; or he could have mentioned his friend Dave who threw his goods off a ship and survived. The hypotheticality of thought experiments is neither here nor in a possible world over there: the relevance of variables is what matters.

The second common criticism of thought experiments that needs deflecting is the tricky issue of intuitions. This vague and slippery term (Williamson 2016, 23-7) usually refers to quick, instinctive responses to a real or hypothetical situation. Thought experiments are often criticized because intuitions are not arguments and can be unreliable (e.g. Jackson 1992, 530; Peijenburg and Atkinson 2003, 306).

I will instead talk of our ‘judgments’ on thought experiments. This includes both quick, instinctive intuitions, and considered, reflective appraisals (Brownlee and Stemplowska, forthcoming section 2.2; Kamm 1992, 9; but compare Kamm 2011, 169). Faced with Nozick’s experience machine scenario, we might respond instantly, or we might use philosophical reasoning. I accept both. Indeed, I have already shown that thought experiments can be used to challenge our intuitions, as with Hobbes, Steiner, Cohen, and Tversky and Kahneman.

2.2. Comparative and non-comparative TEs

I now offer a crucial distinction between comparative and non-comparative thought experiments, depending on whether there is more than one scenario, or just one, respectively. (This builds on the distinction between simple and complex thought experiments – see Brownlee and Stemplowska, forthcoming section 2.3.) Rawls’s original position is non-comparative: it is a
single scenario by which Rawls infers what people would choose under certain conditions of ignorance. Peter Unger’s (1996, 23-6) ‘Vintage Mercedes’ and ‘Charity’ scenarios are comparative: he invites us to judge two scenarios, then compare our judgments:

Vintage Mercedes: Someone has lovingly restored a classic car to mint condition. One day, she finds a trespassing birdwatcher who carelessly cut his leg on barbed wire. The bird-watcher asks if she can drive him 50 miles to the nearest hospital to save his leg. This would cause blood to seep into the seat, costing $5000. She refuses to help.

Charity: A charity asks someone for a $100 donation to save 30 lives. She refuses to help.

Unger suspects that we would feel that the Mercedes owner acted wrongly while the person in the second scenario did not. Yet the cost:benefit ratio is much greater in the former than the latter. (Like Hobbes, Steiner and others, Unger’s thought experiment challenges our intuitions.)

Non-comparative thought experiments can be made comparative. For example, would people in Rawls’s original position choose the same outcomes if they were less risk-averse? This helps us test the importance of risk-averseness in deriving norms. Or, how closely do the original position’s conclusions come to our own society? This helps us test our society (or, perhaps, the original position).

Hobbes’s state of nature can be read non-comparatively or comparatively. Read forwards, from the state of nature to society, the thought experiment is like Rawls’s original position: Hobbes infers, non-comparatively, what rights and duties citizens and sovereigns have, and what the laws of nature are. Read backwards, from society to the state of nature, Hobbes is asking his readers which situation they would prefer – then pointing out that their actions risk a return to a state of nature.

Comparisons can be implicit. Nozick’s initial presentation of the experience machine (1974, 42-4) implicitly invites us to compare a life of guaranteed happiness to an authentic but less happy life. Nozick then makes the scenario comparative by imagining ‘a sequence of machines’ each filling some gap in the earlier machines (1974, 44-5). I am not convinced by these ensuing scenarios, but the principle is right: vary the scenarios to test our norms. This variation is my main focus.
I concentrate henceforth on comparative thought experiments, or on the comparative use of non-comparative thought experiments. Since non-comparative thought experiments can be made comparative, I thus discuss ‘the logic of inference in thought experiments’ as a comparative logic, without denying that some thought experiments work differently.

3. The logic of comparison in thought experiments

Frances Kamm has been the most explicit and rigorous about theorizing comparison in thought experiments. In her Harvard doctoral dissertation, part of which was reprinted in a 1983 essay later included in *Morality, Mortality* (Kamm 2001), Kamm recommends carefully controlled comparisons of two almost identical scenarios. Consider the following situations, to see if killing is worse than letting die:

*Killing*: I drown a child in a bath, to collect his inheritance.

*Letting die*: A child slips in the bath. I could easily save him but do not, to collect his inheritance.

If we judge the cases the same, normatively, we infer that the different feature is normatively irrelevant. If we judge the cases differently, the different feature has affected our judgment (either individually or through an interaction, as I discuss later). Further comparisons would help us test the effect of other variables.

By contrast, consider this flawed comparison:

*Killing*: I drown a child in a bath, to collect his inheritance.

*Letting die in a raging river*: An adult falls into a raging river. I could jump in and maybe save him but do not, to collect his inheritance.

This comparison is flawed: the second example involves much more effort and danger. We could not legitimately infer from this that killing was bad while letting die was not.

These extra variables, effort and danger, can be tested with the following comparisons:

*Killing in a raging river*: I push an adult into a raging river, thereby killing him, to collect his inheritance.
Letting die in a raging river: An adult falls into a raging river. I could jump in and maybe save him but do not, to collect his inheritance.

Most readers, I suspect, will react in the same way as they did for Kamm’s original pairing, above: killing and letting die are morally equivalent, or at least, both are morally reprehensible even if killing is worse. (As I argue shortly, that is so in these scenarios. It may or may not apply more generally.)

These simple examples use controlled comparisons to help us draw normative inferences, both about the conclusion (whether killing is worse than letting die, here) and why.

4. Scientific parallels

More recently Kamm has been explicit that the logic of inference in such thought experiments is like controlled laboratory experiments where we see the effect of changing one variable at a time (Kamm 2011; Kamm 2014). The standard terminology, used throughout this paper, is that the variables we manipulate are the ‘independent variables’ (also known as ‘explanatory variables’) to see their effect on the ‘dependent variable’ (also known as the ‘outcome variable’).

In thought experiments, the dependent variable is our judgment of a scenario, e.g. ‘P is unfair.’ The independent variables are the factors we think may affect our judgment. In Kamm’s drowning child examples, there is almost no effort involved, and no risk, so (she infers) what explains our judgment is the motivation to collect the inheritance.

A few other scholars make similar connections. The most explicit is Gendler (2000, 26), who compares normative thought experiments to Mill’s method of difference, ‘varying the factors which contribute to a situation to see which of them plays which role’ (see also 2000, 115, 136-42, on Mill’s method of agreement). Tamara Horowitz (1998, 367) also makes the connection explicit, although she does not mention variables. Shelly Kagan does not make the parallel explicit but refers to comparing ‘a pair of cases that differ only in terms of the factor in question’ (1988, 5). Kimberley Brownlee and Zofia Stemplowska (forthcoming) talk of ‘variables’ and ‘the hypothesis being tested in the thought experiment (e.g. that turning the trolley is morally permissible/required)’. Elka Brendel (2004, 91-2) states that thought experiments involve ‘the functional dependency of variables by planned and controlled data
change.’ Martin Cohen (2005, 105-6) mentions scientific ideas like replicability, variables and contamination. Keith Dowding (2016, 215, 224-5, 232-3, 238) treats intuitions as ‘data.’ And some scholars discuss framing effects (see below).

So, scientific parallels have been made several times in different ways. But these comments are rare and many insights remain untapped. The first thing to note is that Kamm restricts herself to natural-science experiments, whereas political scientists have a broader range of comparisons. Consider Nozick’s experience machine, which unlike Kamm’s paired scenarios compares two dissimilar scenarios:

*Experience Machine:* I can make a choice to plug myself into a machine that will give me wide-ranging experiences and make me feel happy.

*Real Life:* I lead my own life in my own way. Sometimes I am happy, sometimes I am not. I can experience some wonderful things, but mostly I cannot and do not.

Nozick is asking: would we choose the experience machine or real life? Utilitarians should choose the experience machine, but Nozick thinks we should not do this: such experiences are inauthentic. This suggests that autonomy and authenticity are more important than mere happiness (or that happiness is only valuable if chosen and achieved autonomously and authentically – which is what Mill thought, incidentally, so Nozick’s argument does not trouble all utilitarians).

Thus, Kamm recommends scenarios with just one difference (drowning versus letting drown), but Nozick’s experience machine involves two: having fake experiences and being very happy, versus having real experiences and being less happy. Unger also recommends contrasting comparisons (see the Vintage Mercedes and Charity cases above): paired comparisons ‘must be similar in many ways even while they differ in many others’ (1996, 24).

It appears that Kamm and Unger only see part of the picture: both strategies are legitimate. Kamm and Unger’s logics of inference are somewhat akin to, though not identical to, what political scientists call the ‘Most Similar Systems Design’ (henceforth MSSD) and ‘Most Different Systems Design’ (henceforth MDSD), respectively. (Most political scientists equate these with Mill’s methods of difference and agreement, respectively, but Faure 1994, 316-8 keeps these pairs distinct, producing four not two kinds of comparison, which is effectively my recommendation below.)
The ideal MSSD involves all but one independent variables being the same in the two cases, and a difference in the dependent variable. We then infer that the difference in the independent variable explains the dependent variable’s difference. The ideal MDSD involves only one independent variable being the same in the two cases, but the same dependent variable. If such strikingly different cases end up with the same result, we might infer that it is because of the one identical independent variable.

Political scientists usually say that the MSSD requires a difference in the dependent variable and the MDSD requires the same value of the dependent variable. But thought experimenters should not prejudge the judgment: in Kamm’s cases, one could decide that letting die is bad, or that it is not, say. Here the parallel is closer to natural science experiments where we manipulate an independent variable to see if it does or does not have an effect. Some political scientists indeed recognize that the MSSD and MDSD can apply where dependent variables are the same (Faure 1994, 316-8).

So, we should move beyond the strict logic of the MSSD and MDSD comparisons in themselves as usually understood by political scientists, and see them as examples of broader ways of using controlled comparison to test the importance of variables.

4.1. Differences between thought experiments and scientific comparison

Despite these similarities, important differences remain. Some are relatively unimportant. For example, the dependent variable in thought experiments is usually binary (e.g. ‘is this fair or unfair?’) and only sometimes scalar (e.g. ‘how fair is this?’), but the situation is probably reversed in the social and natural sciences. At the input level, thought experiments do not need randomization, and there are no parallels with things like autocorrelation, analysis of residuals, or out-of-sample testing. At the output level, thought experiments do not produce indices of goodness of fit (like $R^2$ or AIC) or indices of uncertainty (although uncertainty is probably something that political theorists should voice more often – Nozick 1974, xii-xiv, and in history of political thought, Blau 2011).

A far weightier difference is that scientists typically assume that their comparisons aim at uncovering a right answer. But thought experimenters not need assume that a thought experiment will uncover a right normative answer. Some do talk like this (e.g. Horowitz 1998, 367, on using thought experiments to ‘reveal’ norms) but I suspect that two other presumptions are far more common. One is not uncovering norms but uncovering one’s own norms (or helping readers
uncover theirs). As we saw above, manipulating Kamm’s scenarios helps us see why we judge the scenarios as we do. Thought experiments thus help us grasp our ‘tacit commitments’ (Gendler 2013, 3). A second assumption, consistent with the first, is that thought experiments help one achieve reflective equilibrium, i.e. coherence between one’s judgments, principles/arguments, and evidence (Knight, forthcoming). But whereas natural and social scientists typically assume that the results should ideally be identical whoever researches them, thought experimenters should recognize that we often respond differently to the same thought experiment. Perhaps there are ‘multiple reflective equilibria,’ not just one. Thought experimenters often forget this: as Daniel McDermott notes, ‘when a political philosopher mentions “our” intuitions that invariably means “my” intuitions’ (2008, 15).

In discussing ‘scientific’ comparison I am not implying a uniform scientific approach. In that vein, and in the interest of potentially sparking further insights, I should note that I have a ‘regression mindset,’ whereas the kind of causal explanations we use in political theory may be closer to Boolean logic (‘factor f AND factor g BUT NOT factor h UNLESS combined with factor i’). That is what qualitative comparative analysis, or QCA, assumes (Thiem et al. forthcoming, 4-11). Fortunately, this should not affect the key questions of comparison and manipulation discussed above.

5. Internal (and external) validity

The rest of this paper turns on the crucial ideas of internal and external validity. These relate, respectively, to the robustness of the inferences drawn in a given context, and to their generalizability in other contexts (Gerring 2007, 43, 217). Imagine a turnout experiment in a safe Republican district in Utah which finds that telling people the contest is close increases turnout by 2 percentage points more than telling people they have a duty to vote. If the selection is sufficiently random, if the messages are delivered uniformly, and so on, the 2 percentage point result is robust: this is internal validity. But the experiment has little external validity unless we have reason to expect a similar finding in more competitive districts or in other countries, say.

These ideas apply to thought experiments too. For example, if the order in which trolley-problem scenarios are posed affects the outcomes (Otsuka 2008, 109-10), this threatens internal validity: some people’s answers may be a function of question-order effects, not their actual beliefs. (And by the way, we should take seriously the constructivist-inspired possibility that
‘actual beliefs’ is a misnomer, that every answer has question-order effects and framing effects, that there are no neutral beliefs to be uncovered by well-posed thought experiments. This should not stop us doing thought experiments, just as it does not stop social scientists doing surveys. It just means we should all be cautious with our data.) Meanwhile, if our answers to trolley problems do not apply to dirty-hands problems in political theory, say, this is a problem of external validity.

Internal validity can be threatened in two ways: judging a scenario wrongly, and judging a scenario correctly but getting the wrong causal explanation of their judgment. To distinguish these two, consider a social-science idea which is widely recognized in discussions of thought experiments: ‘framing effects’ (Rachels 1975, 80; Nozick 1993, 60; Horowitz 1998, 369-81; Brownlee and Stemplowska forthcoming, section 3.2.2). Imagine that we are probing the legitimacy of military intervention, using thought experiments involving US intervention in fictional countries. But this will bias some readers because of their views on previous US military interventions. Some readers will shake off these feelings, others will not. If they oppose military intervention in this biased scenario when a more neutral case would have led them to support it, say, they have judged the scenario wrongly. If they oppose it and would have opposed it with a more neutral case, but if they attribute this opposition to the wrong factor because the US frame affects their reasoning, then the second threat to internal validity has happened: although they have judged the scenario correctly, their causal explanation is wrong. This is one reason, note, why thought experimenters often choose very ‘distant’ examples, to avoid framing effects from ‘real’ examples.

5.1. Increasing internal validity by controlled comparison

The key to internal validity is controlled comparison – a ‘theoretically informed combination of control and variation’ (Slater and Ziblatt 2013, 1308). Control is far easier for thought experimenters than for most social scientists, since we can imagine highly stripped-down scenarios which control for most factors, whereas social scientists cannot simply give two countries the same level of education, say. (Replicability, interestingly, is also easier for political theorists than most natural and social scientists.) My main focus in what follows is thus variation.

I will describe four ways of manipulating independent variables (although they are actually different ways of saying the same thing). First, we can change the value of an
independent variable. If trolley problems make you conclude that we should not intentionally kill
one person to save five, would you judge differently if you could save 5000? Or 5 billion? (See
Thomson 1990, 167.) Manipulating variables like this can help us test our judgements.

This is equivalent to what social scientists call ‘sensitivity analysis,’ seeing how altering
the value of an independent variable affects the dependent variable (which in a thought
experiment is our judgment). Social scientists usually do sensitivity analysis to find the size of
the effect (e.g. a 1 percentage point shift in votes leads to a 3 percentage points shift in seats). In
the trolley-problem example just mentioned, sensitivity analysis shows when, if ever, the
dependent variable flips, from wrong to acceptable, say. (This number will vary from person to
person. It will rarely be a clear point (Parfit 2011, 229), for Sorites-related reasons.

Second, we can remove one or more independent variables. In Unger’s Vintage Mercedes
scenario, would our judgement change if the birdwatcher were not careless? This is effectively
the same as the first type of manipulation: the degree of carelessness is high in one scenario, low
in another.

Third, we can add new independent variables. In Singer’s (1972, 231-2) discussion of
whether we should help a child drowning in a nearby pond, would our judgement change if other
people were present? (This is like a natural-science experiment where one introduces extra
variables into the experiment, e.g. will higher oxygen levels affect plant growth?) This too is
effectively the same as the first type of manipulation: the number of onlookers is zero in one
scenario, positive in another.

Fourth, we can check for interaction between our variables. Technically, this is
equivalent to adding a new variable (the variable of ‘interaction between existing variable P and
existing variable Q) but it helps to think of this as a separate technique. I will examine this more
in section 5.3.

5.2. Omitted variable bias

I now address threats to internal validity. The first threat, omitted variable bias (OVB),
means that at least one independent variable is not included in the causal model even though it
does actually have a causal effect. Imagine that a social scientist wants to explain children’s
height. Unsurprisingly, age is the leading influence. But the social scientist forgets to check for
gender. This is an omitted variable (a ‘confounder’) in the regression, because boys tend to be
taller than girls: so, although gender does actually influence height, it was left out of the
regression. Omitted variables can bias the results, leading to faulty causal inferences. Omitted variables can be hard to spot, or hard to measure, and ‘virtually all social science analyses are plagued by omitted variable bias to one degree or another’ (Jargowsky 2005, 924).

Kamm’s first drowning-baby comparison looks like OVB. Assume, for the sake of argument, that we conclude that in this scenario, killing is as bad as letting die. I still do not think Kamm is right that the reason is because of killing versus letting die in itself (see Kamm 2001, 31, 33; also Rachels 1975, 79-80). I am not sure that this position is meaningful: killing is always of someone, by someone, in a certain context. But leaving that aside, Kamm does not mention a factor which may be doing some of the work here: most of us would say that a capable adult has a moral responsibility to help an unaccompanied child in difficulty if both are in the same location and no one else is. We can test this by manipulating Kamm’s scenarios to make this issue more prominent:

**Assassin:** I pay an assassin $50 to drown a child in a bath in a different country, 5000 miles away.

**Letting die abroad:** I read that a child is often left unattended in a bath in a different country, 5000 miles away. I could pay someone $50 to look after him but I do not.

Readers’ responses to this will vary, I suspect. For some, $50 is a small price to pay to save a life, so one has a responsibility to help: call these respondents ‘Assisters.’ For others, whoever keeps leaving the child in the bath has that responsibility: call these respondents ‘Decliners.’ Assisters and Decliners agree that we have a moral responsibility to help unaccompanied children but disagree about when. But both presumably agree that these situations include Kamm’s original scenarios. Kamm’s normative judgment of these scenarios seems correct but her explanation does not: OVB leads her to claim that the cause is killing/letting due ‘in itself’.

In Kamm’s initial scenarios, OVB arises from a variable that is not mentioned in the thought experiment but which is actually doing some of the work in the background without us noticing. This kind of thing is probably extremely common.

OVB can also arise from a variable that is mentioned in the thought experiment but which we forget may be influencing us. For example, Goodin (1982, 10) notes that Thomson’s famous violinist thought experiment starts with the violinist being kidnapped. Thomson herself admits that this might influence us to think it legitimate to unhook oneself from the violinist
Now imagine for the sake of argument that we repeat the thought experiment and still conclude that it is legitimate to unhook oneself. If so, we can be more confident about why we feel this way: kidnapping is no longer doing any work in the background. (This assumes that we can rerun the thought experiment and forget about the kidnapping. But see Otsuka 2008, 109-10 on hangover effects from previous thought experiments. Like social scientists, we must consider the reliability of our data.)

I doubt that OVB is as likely for thought experimenters as for social scientists, and what is OVB for one person may not be for another, since we respond to many thought experiments differently. But OVB is probably more widespread than we think, and while it is nothing to be ashamed of – other people can easily spot factors we have overlooked – we should nonetheless be awake to it.

5.3. Interaction effects

One form of OVB which needs special attention is interaction effects. Imagine that two variables have no influence unless both are present. What has an effect is not the first variable, or the second variable, but the interaction of the two. They do not, strictly speaking, need to ‘interact’: an interaction effect can be like a rabbit that runs away not when it sees a fox, or an eagle, but only when it sees both together, even if the fox and eagle are unaware of each other.

In social science, it is more common for one or both variables to have their own effects and for there to be an added interaction effect. Perhaps old people are in general more likely to vote than young people, and women are in general more likely to vote than men, but old women are even more likely to vote, on top of the previous two effects. Interaction effects are common in social science, and can be simple to model: in the above example, we might include three variables – age, gender (a ‘dummy variable’, e.g. 1 for women and 0 for men), and age-multiplied-by-gender. Our statistical software then tells us the size and significance of each variable.¹

In thought experiments, interaction effects are recognized by Michael Stocker (1987, 277) when discussing Kantian arguments which show that virtues like courage or friendship are not intrinsically good, since courageous acts of evil, or helping a friend commit fraud, are not good. What Stocker describes as ‘the complex of f-and-not-g’ is what I describe as an interaction effect.

¹ And if you use SPSS, your friends will tell you to use STATA or R.
Kamm initially discusses the effect of killing or letting die ‘in themselves,’ but does recognize interaction effects later (2001, 55-6), e.g. when killing and letting die depend on the ease of effort. As she notes, if we suspect an interaction, ‘we can simply create another set of comparable cases without this contextual factor, and see whether our judgment about these new cases differs’ (2001, 56). Interaction effects can be tested like any other variable, but if they are normatively relevant, excluding them from a causal explanation is a form of OVB.

Interaction effects are potentially very troubling for thought experimenters. Goodin criticises excessively ‘clean’ thought experiments which ‘obscure those interactions that are so typical of the complex cases in the real world.’ For example, retributivists attack utilitarians using the thought experiment of a mob seeking to hang an innocent person, utilitarians reply with the thought experiment of a criminal whose punishment would have no effect on the crime rate.

Since each case seems compelling, the conclusion appears to be that neither deterrence nor retribution is adequate justification for punishment. Actually we have proven that both are, but only in conjunction. Guilt and deterrence are individually necessary and jointly sufficient conditions of morally justified punishment. But since the justification depends on the interaction between the two, dropping each condition one at a time and asking ‘what would you say then?’ naturally leads us to mistaken conclusions (Goodin 1982, 9).

Strictly speaking, we have not ‘proven’ anything: proof is for deductive inferences only, and these normative thought experiments are inductive. But Goodin’s objection is right: these thought experiments generate faulty inferences if we treat causality as non-interactive. The problem, though, is not thought experiments as such, but misconceived thought experiments where we have not controlled and manipulated to test interactions. Goodin’s argument suggests exactly these tests. And note that ‘real’ examples can lead to identical errors.

5.4. Spurious correlation

Spurious correlation is the reverse of OVB. OVB means that a variable is actually causally relevant but has been excluded from the analyst’s causal explanation e.g. the regression model. In thought experiments, omitted variables actually influence one’s judgment but are not spotted: one’s judgment on the scenario is correct, but one attributes this to something else.
The reverse situation, where a variable is not actually causally relevant but is included in the causal explanation, is ‘spurious correlation.’ (This is different to what social scientists call ‘superfluous’ or ‘irrelevant’ variables, which are not talked of as causing bias: if irrelevant variables are not correlated with the dependent variable, they merely increase the standard errors of the coefficients, making the coefficients of the causally relevant independent variables less precise – Gujarati 2012, 121.) Again, spurious correlation means that one’s judgment of a scenario is right but one’s causal explanation is wrong.

Consider the brilliant and very funny internet video comparing rape to forcing someone to drink tea (https://www.youtube.com/watch?v=oQbei5JGiT8). It shows someone repeatedly trying to force someone to drink tea – a ‘completely ludicrous’ situation, it rightly says – and ends by noting that you wouldn’t want to force someone to drink tea against their will, so why force someone to have sex against their will?

Strictly speaking, though, this well-intentioned comparison is flawed. To desperately want to give someone tea is odd; to do so against their will is ludicrous. But many people do desperately want to have sex with someone; to do so against their will is abhorrent and appalling, but it is not ludicrous in the same way as the tea example. Most people watching the video will agree that we should not force someone to drink tea against their will, and that we should not have sex with someone against their will, so the judgment of the thought experiment is right: rape is wrong. But the thought experiment seems wrongly designed to make us think that one reason not to do either thing is the ludicrousness of the situation. However, only the tea scenario is ludicrous. The real factor driving both situations is still consent (as the video’s title makes clear): no means no, whether or not you want someone to say yes.

In this thought experiment, ludicrousness is what we might call an ‘excess variable,’ the opposite of an omitted variable. It is not what drives our judgment that non-consensual sex is wrong, but the thought experiment is posed so cleverly that this is an easy mistake to make. I suspect this problem is very common in thought experiments.

5.5. Testable implications

A key scientific tool is ‘observable implications’ or ‘testable implications.’ (The latter term is more apposite for us.) If $P$ is right, what would we expect to find, and what would we not expect to find? We all know this idea, but making it explicit helps. Note, incidentally, that testable implications are not best seen as part of ‘hypothetico-deductive’ analysis, as orthodox
social scientists imply (e.g. Elster 2007, 17–20, 52–66, 246–56). ‘Hypothetico-inductive’ is usually more apt: the implications are merely what we might well expect to see (Blau 2015a, 1187).

Consider this example. Suppose we worry that Kamm’s drowning child scenarios are subconsciously influencing our normative judgments because of the use of children. A testable implication is that we might judge the scenarios differently if they involved adults. Since an adult cannot easily drown or be drowned in a bath, we could rerun the scenario in the following way:

*Electrocution.* I rewire a light switch which I expect a relative to turn on, hence fatally electrocuting him, to collect his inheritance.

*Passive Electrocution:* I see that a light switch could give someone a fatal electric shock but do not stop a relative from turning it on, hence letting him be fatally electrocuted, to collect his inheritance.

For me, and presumably most readers, these two scenarios elicit the same reaction as with Kamm’s original case. This suggests that our initial worry was wrong: using children in the scenarios is not influencing us.

Testable implications are a powerful tool for thought experimenters, helping us rerun thought experiments to test the effect of different variables.

6. External validity

Having discussed internal validity, i.e. whether we have drawn the correct inferences about causes and effects in thought experiments, I now turn to external validity, i.e. whether we can extrapolate our judgments from thought experiments in other situations, especially real-life ones. How easily can we apply trolley problems to ‘dirty hands’ problems in politics? Do Kamm’s drowning child scenarios apply to letting people die in developing countries?

To understand this we need to distance ourselves somewhat from orthodox social-science accounts of external validity which hold that it ‘rests upon the representativeness of the sample’ (Gerring 2007, 217; emphasis added). Representativeness has even been described as a ‘fundamental’ precept from which ‘[n]o reasonable definition of internal and external validity can depart’ (Slater and Ziblatt 2013, 1305).
In fact, this claim is absent from some accounts (e.g. Mark and Reichardt 2001; McDonald 2005, 939, 946). I would say – and this is going to be very important for my analysis of thought experiments – that representativeness is a secondary quality, deriving from a logically prior issue: relevance of variables. We call a sample representative if relevant variables are representative even if irrelevant ones are not. Consider phone-in polls, which typically attract above-average proportions of people who are angry about the issue being polled. This sample would presumably be representative in terms of hair colour but not ideology, but ideology is the relevant variable, so we call the sample unrepresentative and pronounce its external validity weak.

Moreover, thought experimenters need not worry if a thought experiment is representative; I do not even know what that means in cases like Nozick’s experience machine. The real issue is whether the variables are sufficiently relevant for us to apply a thought experiment in other contexts. Kamm’s drowning child scenarios lack many real-life variables that hinder its external validity, e.g. when thinking about letting people die of starvation in other countries, where others also have responsibility for them. (Singer would disagree: other people having responsibility is not a relevant variable. All that matters is whether you can help another with ease.) With Nozick, the variables do seem to be relevant to any claim about the inherent correctness of some kinds of utilitarianism: this is why thought experiments can be powerful tools for probing abstract normative principles. The variables are not relevant in every instance, of course, e.g. we might recommend the experience machine to someone in a concentration camp (Schmidt 2002, 211).

So, representativeness is not the main issue for thought experimenters: relevance of variables is key. I suspect this is true for social scientists too.

I will address two issues of external validity. The most important is parsimony, but first I discuss counternomic scenarios. In both cases, I do not mean to imply that external validity problems must or could be fully overcome. After all, social scientists do not need perfectly specified models to give sensible action-guiding advice in new contexts. Nor do thought experimenters need to plug in every variable before giving advice, just as many normative principles leave out relevant variables: this is why laws are regularly applied in contexts that lawmakers could not consider. But we should be aware that the internal/external validity tradeoff in social science has a counterpart in political theory: thought experiments help us filter out details and test the effect of some key variables, but risk filtering out others.
In the ensuing discussion, what matters most is external validity: we should try not to be overly skeptical of thought experiments that use ‘wacky’ examples (Brownlee and Stemplowska forthcoming, section 4). But I do think that thought experimenters sometimes make life hard for themselves by using examples that are unnecessarily weird, or just very clumsy; an example of the latter is Kamm’s (2001, 71-2) plane/lever example. Such scenarios invite readers to wonder: if you could only make your case this way, are the lessons really of general applicability? As we will see, perhaps the answer is ‘yes.’ But there must be a concern that just as hard cases make bad law, perhaps wacky cases make bad norms.

6.1. Counternomic examples

Some writers criticize thought experiments which use what Onora O’Neill calls ‘counternomic’ examples that break physical laws, like Thomson’s man who can save thousands of cows from horrible suffering simply by being punched, and the violinist case (O’Neill 1986, 20-1). Counternomic thought experiments do not have direct scientific parallels, obviously.

Derek Parfit replies that such cases ‘may provide a partial test for our moral principles. We cannot simply ignore imagined cases’ (1984, 388; emphasis added). That strikes me as appropriately balanced: disallowing counternomic thought experiments is too strong, but caution is needed. Simply rejecting counternomic scenarios is no more convincing than supporting them regardless. The logic of comparison requires critics and defenders to ask if counternomic scenarios lack relevant real-life variables that might change our judgments, and if variables in the hypothetical scenarios are relevantly missing in real life – as noted by Goodin (1982, 10). The primary focus should be on variables and outcomes, not their counternomic nature. After all, I do not have a right to wave a magic wand that would cause O’Neill extreme pain and then kill her while making some cows and violinists happier: that normative inference is legitimate even though my example is counternomic, because the relevant variables are the same.

6.2. Parsimony

The more important issue of external validity is parsimony. To avoid equating parsimony with simplicity, I define parsimony as exclusion of relevant variables. (See my earlier discussion of Foot’s craniotomy thought experiment, a simple scenario which successfully captures the relevant variables in many cases.)

Social scientists often use parsimonious models/explanations which intentionally leave out relevant variables. There are several reasons to do this, including easier formal modelling,
and for philosophical reasons like Occam’s razor. There are also good technical reasons for parsimony (Schrodt 2014, 288).

But parsimony is not always a virtue: we may want better specified models if we want to give ‘guidelines for social practice’ (Przeworski and Teune 1970, 22-3). And guidelines for social practice are precisely what many political theorists seek. We should be careful when using parsimonious thought experiments to generate action-guiding principles.

So, OVB and parsimony involve different problems. With OVB, the dependent variable is correctly predicted but the model is incorrectly specified. So, the explanation is not quite right – a problem of internal validity. For example, Kamm’s first comparison generates correct judgments, but her explanation is wrong: moral responsibility does some of the work.

Parsimony can have the same effect, but may also mean that the dependent variable is incorrectly predicted. This is an external validity problem: if we extrapolate from the parsimonious model to real-world situations, our judgment may need to change. For example, even if we are right that letting die is morally blameworthy in Kamm’s scenarios, it may not be in some real-life situations.

Social scientists often describe a tradeoff between internal and external validity: the more closely controlled a comparison, the harder it is to apply in other contexts (e.g. Gerring 2007, 43). But we can partly overcome these problems with enough knowledge of the relevant variables, of course.

This has three important corollaries in debates over parsimonious thought experiments. The first is that parsimony-based objections ideally need some discussion of why variables might differ in other contexts. By contrast, O’Neill (1986, 20-1) is too quick with her critique of ‘socially decontextualized’ thought experiments, like Nozick’s (1974, 34-5) ‘Ray Gun’ scenario:

Ray Gun: You are at the bottom of a deep well. Someone throws a live person at you. May you use a ray gun to disintegrate the falling body before it hits you?

Nozick has, albeit awkwardly and too implicitly, envisioned a scenario with two choices: kill an innocent person and live, or refuse to kill and die (presumably with the innocent person surviving, here by landing on you). O’Neill’s criticism implies that adding real-life variables makes one less likely to support killing an innocent person. I can imagine a few such cases, e.g. if it was my fault that I was down the well. But I suspect that in similar choices as stark as
Nozick’s, many of us would accept that it is legitimate (albeit horrendous) to kill an innocent person and live, whether or not we would do so.

So, Nozick’s scenario probably works fairly well. It filters out irrelevant variables, and in these situations most variables are irrelevant: on my definition of parsimony, Nozick’s scenario is not very parsimonious. Without showing what social context adds, O’Neill’s critique should not trouble Nozick. Perhaps other critiques will fare better (see e.g. the critique of just-war thought experiments by Dunford and Neu 2016).

The second, related corollary is that parsimony-based critiques should ideally consider whether the tradeoff between internal and external validity can be partly overcome by seeing if our judgements change as we make thought experiments more realistic. Another critique which does not quite go far enough is Goodin’s objection to thought experiments which are ‘too stripped down to be of any real policy guidance’ (1982, 8). His example is Nozick’s Wilt Chamberlain thought experiment: if people freely pay to watch Wilt Chamberlain play basketball, Nozick argues, it does not seem objectionable if he ends up richer than everyone else (1974, 161-4). Nozick uses this to oppose government redistribution of free-market transfers, but if we accept Chamberlain’s wealth, Goodin replies, it is ‘merely because we have not considered all the things that he might do with it,’ such as buying slaves or hoarding all of the food in the area (Goodin 1982, 8-9). This is not quite fair to Nozick, who does not allow slavery unless it is self-chosen (1974, 331), and who limits how much of a resource like food any individual could buy (1974, 178-82). But for us, the point is that Goodin’s criticism actually calls initially for a revision, not (yet) a rejection, of Nozick’s thought experiment: let us first re-run it in different ways and see if our conclusions change (Cohen 1995, 23).

The third and ultimately most important corollary is to get a better sense of how big our external validity problems are. A scientific parallel again helps. Social scientists know that changing their regression specifications can sometimes radically affect their causal inferences, but this is likelier with poor data and questionable statistical analyses (Treisman 2007), or the wrong statistical assumptions, where ‘even minor changes in model specification can lead to coefficient estimates that bounce around like a box full of gerbils on methamphetamines’ (Schrodt 2014, 289).

Do thought experiments have a similar methamphetamine problem? We may find it hard to say without more assessments of the external validity of parsimonious thought experiments, e.g. seeing if our judgments change as we gradually move from parsimonious to realistic
scenarios. But that test is hard, given the number of relevant variables and the difficulties of rigorously testing even a single one. Christopher Achen, a quantitative political scientist who fiercely criticizes widespread statistical practises, recommends a ‘rule of three’ – no more than three independent variables, for better testing of models (2002, 446-8). Do thought experimenters need something similar? Again, I suspect it is hard to say. A more precise methodological language and more focused methodological discussions should help.

7. Conclusions

This paper’s main contributions are about testing and strengthening thought experiments. I have also sought to cast new light on longstanding debates about thought experiments, deflecting some criticisms while suggesting that other criticisms are quite troubling for thought experimenters.

The parallels I draw between science and political theory also inform a prominent issue in the philosophy of science and philosophy of social science: ‘naturalism.’ One common understanding of this term – there are other understandings – is that there is ‘continuity’ between philosophy and the empirical sciences (e.g. Ritchie 2008, 195).

Yet philosophers are quite vague about what ‘continuity’ entails. The similarities highlighted earlier are thus instructive. Continuity does not mean identity, of course, and I am not implying that there is a single scientific method. Nor do social-science ideas all derive from the natural sciences: we should proud of political-science developments like process-tracing and fuzzy-set QCA, for example. Thought experimenters might conceivably generate methodological insights for natural and social scientists.

Interestingly, there are different parallels between science and the history of political thought (and history of philosophy), as I implicitly or explicitly note in five publications (Blau 2011; Blau 2012; Blau 2015a, where the parallels are implicit but clearest; Blau 2015b, where the parallels are most explicit – see pp. 42-50; and Blau forthcoming). Scientific ideas like hypothesis-testing and triangulation, I argue, help more than the literature on hermeneutics. Box-fitting classification – ‘schools of thought’ like contextualism and Straussianism – accentuates diversity without addressing some unifying methodological principles. But these publications of mine say little about comparison, and largely avoid normative political theory.
Why do these parallels between science and political theory matter? Andrew Rehfeld notes political theory’s decline in many US universities, partly because ‘the discipline [of political science] struggles to understand why theorists belong to it’ (Rehfeld 2010, 465-7). He responds by showing methodological similarities between science and political theory. But here we diverge. Rehfeld largely reduces scientific methodology to falsification (Rehfeld 2010, 473-4), which I ignore above. (For my unorthodox views on falsification, see Blau 2015b, 1186; see also Dowding 2016, 104, 107, 113-27.) Rehfeld says little about scientific logics of inference. Note too that he places political theory within political science (see especially Rehfeld 2010, 475-6), whereas my ‘continuity’ approach merely highlights similarities.

I hope that highlighting these similarities might also make politics departments happier places to work in. I have heard political scientists and political theorists make dismissive comments about each other’s work (and I used to make such comments myself – mea culpa). Political theorists’ disdain sometimes centres on caricatures of political science: number-crunching, law-making, and so on. Political scientists’ disdain can actually be more insulting. ‘All the problems of politics are empirical,’ a British professor once barked at Martha Nussbaum, stabbing his finger at her. A political scientist who I particularly admire once said ‘I don’t know why anyone studies political theory.’ I hope such comments might be less common if we knew more about each other’s methodology. It would also help if political science methodology courses could occasionally make links to political theory.

It seems fitting to end by paraphrasing one of the greatest contemporary exponents of thought experiments, Derek Parfit. His monumental On What Matters, originally called Climbing the Mountain, uses thought experiments and other normative techniques to argue that Kantians, contractualists and consequentialists are far closer than they think: they are ‘climbing the same mountain on different sides’ (2011, 419). The mountain metaphor applies differently here: political theorists and political scientists are often climbing different mountains, but sometimes we use the same tools.
REFERENCES


Peijenburg, Jeanne, and David Atkinson, 2003. When are thought experiments poor ones?, *Journal for General Philosophy of Science* 34, 305-22.


Thiem, Alrik, Michael Baumgartner, and Damien Bol, forthcoming. Still Lost in Translation! A Correction of Three Misunderstandings Between Configurational Comparativists and Regressional Analysts, *Comparative Political Studies*. Online first version.


Treisman, Daniel. 2007. What have we learned about the causes of corruption from ten years of cross-national empirical research?, *Annual Review of Political Science* 10, 211-44.


