

Manuscript version: Author's Accepted Manuscript

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:

<http://wrap.warwick.ac.uk/142244>

How to cite:

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

Directed Reflective Equilibrium: Thought Experiments and How to Use Them

A widely adopted method in philosophy is reflective equilibrium (hereafter RE).¹ According to this method, philosophers should aim to construct a theory that maximally coheres with considered moral judgments and general principles as well as a wider range of beliefs and facts.² The theorist works back and forth between these commitments, discarding previous beliefs if necessary, to reach an equilibrium. A central component in the method of RE is the use of imaginary and real-world examples, thought experiments and intuition pumps. For simplicity, let us call these real or imagined realities *cases*.

The use of cases in normative theorising has a long and illustrious history but has also been subject to a number of criticisms, which, in turn, threaten the validity of the method of reflective equilibrium. First, intuitions about cases are vulnerable to debunking or manipulation. For example, we are more averse to prospective losses than we are attracted to prospective gains.³ If the framing of cases can affect our intuitions about them, how can intuition be a reliable guide to moral principle? Second, cases often simplify and abstract from real world situations. Some worry that intuitions about fantastical cases warp our sense of morality; or that they encourage our moral thinking to become unrepresentative of or detached from real-world crises.⁴ A suspicion of abstractionism underpins much historical scepticism towards moral theory in general,⁵ and similar worries can be raised about hypothetical cases. Third, coherentist approaches to methodology (like RE) face a general challenge about what to do in the event of inconsistency between our intuitions or between intuitions and basic principles.⁶ To be sure, theorists have developed resources to help overcome this impasse: the robustness of judgments, the vulnerability of intuitions to debunking, theoretical parsimony, and so on. However, what

¹ See John Rawls, *A Theory of Justice*, (1971), 20 for the introduction of the terminology.

² See Rawls, *A Theory of Justice* and Norman Daniels, *Justice and Justification: Reflective Equilibrium in Theory and Practice* (Cambridge: Cambridge University Press, 1996), Ch. 1.

³ Daniel Kahneman, *Tanner Lectures in Human Values*, (University of Michigan, Ann Arbor, 1994)

⁴ Allen Wood, 'Humanity as an End in Itself' in Derek Parfit, *On What Matters*, Volume 2, (Oxford: Oxford University Press, 2011) and Mathias Thaler, 'Unhinged Frames: Assessing Thought Experiments in Normative Political Theory', *British Journal of Political Science* 48 (2016), pp. 1119–1141.

⁵ Onora O'Neill, 'Abstraction Idealization and Ideology in Ethics', *Royal Institute of Philosophy Supplements* 22 (1987), pp. 55–69.

⁶ For similar queries about reflective equilibrium, see J. Arras, 'The Way We Reason Now: Reflective Equilibrium in Bioethics' in *The Oxford Handbook of Bioethics*, B. Steinbock (ed.) (New York: Oxford University Press, 2007), pp. 46–71 and T. Kelly and S. McGrath, 'Is Reflective Equilibrium Enough?' *Philosophical Perspectives*, 24(1) (2010), pp. 325–359.

RE is still lacking is a sense of how cases ought to be ordered in theoretical enquiry, given their different uses.

In this paper we defend a revised version of RE that we call Directed Reflective Equilibrium (hereafter DRE). DRE, like its predecessor, accepts that neither intuitions nor basic principles are immune to revision and that our commitments on various levels of philosophical enquiry should be brought into equilibrium. However, it also offers guidance about how different types of cases ought to be used, thus exhibiting cases at their best and overcoming some of the methodological shortcomings faced by RE. With a clearer typology of cases in mind, an order of their usage suggests itself which helps overcome the pitfalls of RE.

The suggested order of the DRE proceeds as follows: First, philosophers should start from what we call “seed cases”. Seed cases are situations or dilemmas, usually from real life, that capture our moral attention and elicit strong, if unsystematized, intuitions. Second, these cases are then “decomposed” into the relevant moral factors at play in the case. Doing so allows the philosopher to construct intuition-generating “controlled cases” that represent a class of issues in which the relevant factor is present. Testing different versions of these cases against each other, the philosopher then seeks to “organize” the general virtues and the strength and scope of the elicited intuitions into principles. As in standard RE, this organization will require going back and forth between principles and concrete judgements in representative cases. Third, to further test these principles, philosophers can create “argument cases” that elicit the recognition of reasons rather than intuitions, seeking to support principles on the one hand, and challenge biases, metaphysical beliefs, and underlying conceptual assumptions that may taint our intuitions on the other. Fourth, principles that cohere with both intuitions and reasons can be “veiled” in the final type of cases: “construction cases”, which set up choice situations into which these fundamental principles are already incorporated, making choices that do not accord with the principle impossible. Figure 1 shows the two-by-two case taxonomy we are proposing, with the two dimensions indicated at the top and left of the figure. The four types of cases we distinguish all have a role to play in DRE, ideally in the order indicated by the grey arrows, starting with seed cases. And because principles play a central role in RE, all but the seed cases have a function related to the formulation, testing, support and systematization of principles.

Various stages of our model will be familiar to many philosophers. Individuals, and philosophical debates more broadly, often employ cases in the ways we recommend. Our purpose here is not to fundamentally challenge the way cases are currently used, or to suggest a radically different usage, but to systematize pre-existing elements of best practice and to highlight the advantages of a specifically directional approach. In the next section, we develop the model

in greater detail, and in section XXXXX, we argue that the model improves on RE by addressing some of the pitfalls of the case-based methodology mentioned above.

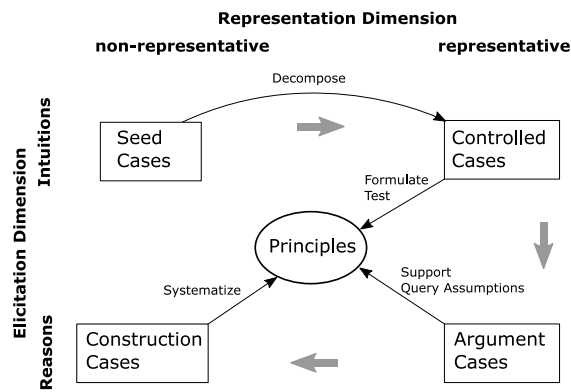


Figure 1: Directed Reflective Equilibrium Case Use.

Seed Cases

The first stage of our model employs what we call seed cases. These are cases that capture the moral phenomenon we wish to investigate, without making any initial effort to decide what factors are morally salient, or to separate relevant from irrelevant factors. Many debates in moral philosophy have been inspired by real cases that seem to capture something important about the normative landscape. For example, decisions in war may inspire discussions of principles in just war theory,⁷ cases of famine or other humanitarian crises may inspire discussion of the duty of rescue,⁸ acts of terrorism and political torture may prompt discussions of harming others as a means to an end. As well as being taken from real scenarios, good seed cases also frequently elicit strong, but conflicting or conflated intuitions. Cases of famine, for example, may raise complex moral problems involving, among other things, the distinction between positive and negative duties, the stringency of assistive duties, historical and contemporaneous responsibility of wealthy countries for poverty-related hardship, and many more. The same is true of harming in war, terrorism, and many other common seed cases in moral philosophy. These cases often capture important, but complex moral problems. They pull our intuitions in different directions, perhaps in accordance with

⁷ For example, bomber.

⁸ Singer. Gerver.

pre-existing moral or political sensibilities, and almost always involve a complex intersecting of different morally salient facts.

From Seed to Controlled Cases: Decomposition

Seed cases provide a starting point for theoretical enquiry, but their complexity or “murkiness” can be problematic. The purpose of the next stage, decomposition, is to identify a range of factors that have potential moral salience and extract them from the seed case. Once we have extracted as many of these factors as possible, they can be formulated into their own cases and thereby separated from factors with which they are coexist in the seed case. Let’s take the example of harming in war to demonstrate this process. Suppose we take as our seed case a report of a soldier killing an unarmed combatant in war. We then break the case down into a list of factors that might have moral salience. There may be many such factors, including: (1) orders within a military hierarchy, (2) the chaotic context of war, (3) epistemic uncertainty, (4) the status of the victim (combatant or non-combatant) (5) whether the victim was armed, (6) the culpability of the decision to kill, (7) whether wrongdoing was foreseeable, (8) the moral significance of causation, and perhaps more. Each of these factors can then provide the basis of further cases where they are separated from others. Many revisionists in just war theory, for example, comparing situations in war to structurally similar cases of interpersonal harm, to isolate relevant factors from, say, the chaotic context of war or the epistemic uncertainty that pervades decisions in war.⁹

Controlled Cases

Controlled Cases for Separating Factors

Building on the output of the decomposition, philosophers can systematically integrate the different factors into different hypothetical or constructed cases. In our previous example we saw how we might separate factors like culpability and causation, consider self-defence outside the context of war entirely, stipulate epistemic certainty, and so on. Such cases are made possible through decomposition by separating and isolating the different normative factors at play in a seed case. We will refer to cases used in this stage of the process of DRE as *controlled cases* to

⁹ McMahan, Killing in War.

emphasise their use in separating factors.¹⁰ Unlike seed cases, which are singular, however, these cases aim to *represent* a particular factor that is present many in real life situations.

We can clarify the representative dimension of cases by borrowing from the discussion of models in the philosophy of science. Put in a nutshell, a model is a representation of a target system, and the relevant relation between target system and model is a similarity relation. The basic idea goes back to Ronald Giere¹¹ and was developed in detail by Peter Godfrey-Smith, Michael Weisberg and many others.¹²

We can think of most models as structures that are relevantly similar to their respective target systems. For example, the drawing of a cell in a biology textbook is relevantly similar to many different cells in the real world. It is an idealized exemplar of real cells.¹³ What makes it similar is that certain structural features are alike. That is true even though no real cell might look precisely like the drawing (quite apart from the fact that real cells are not made of ink and paper, have three rather than two dimensions, and so on).

One popular approach in the sciences has a “hub-and-spoke” structure, as Godfrey-Smith points out:

“In these cases, what scientists do is give an exact description of one case of the target phenomenon, which acts as a “hub” that anchors a large number of other cases. The “other” cases include all the actual-world ones; the hub is a fiction. The central models of both evolutionary change and population growth within modern biology work like this, for example.”¹⁴

This should sound familiar to the theorist drawing on paradigmatic thought experiments. Consider Peter Singer’s case of the child drowning in a shallow pond that you pass on your way to work and who you could save at little cost to yourself.¹⁵ This is a fiction, but many structurally

¹⁰ A similar, but more minimalistic way of depicting this use of thought experiments (termed “heuristic thought experiments”) can be found in Brun, G. (2017). Thought experiments in ethics. In *The Routledge Companion to Thought Experiments* (pp. 195-210). Routledge.

¹¹ Ronald Giere, *Explaining Science: A Cognitive Approach*, (Chicago: University of Chicago Press, 1988).

¹² Peter Godfrey-Smith, Peter, ‘The Strategy of Model-Based Science’, *Biology and Philosophy* 21 (2006), pp. 725–40; ‘Models and Fictions in Science’, *Philosophical Studies* 143 (1) (2009), pp. 101–16 and Michael Weisberg, *Simulation and Similarity: Using Models to Understand the World*, (Oxford and New York: Oxford University Press, 2013).

¹³ Weisberg, *Simulation and Similarity*, p. 18 and Stephen Downes, ‘The Importance of Models in Theorizing: A Deflationary Semantic View’, *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1 (January) (1992), pp. 142–53.

¹⁴ Godfrey-Smith, ‘Models and Fictions in Science’, p. 107.

¹⁵ Peter Singer, ‘Famine, Affluence, and Morality’, *Philosophy & Public Affairs*, 1(3) (Spring 1972).

similar cases are “anchored” around this “hub” case. The “hub”, or core normative feature, of the case is the fact that you can provide great benefits to others at trivial cost to yourself. This feature is present in many social, legal, and political dilemmas, where it coexists with and is complicated by many other factors.

The sciences prefer models as equations, computer code or scale models. In normative theory, however, most models are “word models,” stated purely in narrative form. The occasional formalized game-theoretical model can be found but remains the exception rather than the norm. However, this should not detract from the fact that the function is very similar: to represent a target system in a way that makes it more amenable to analysis than the real-world cases it represents. In physics, frictionless planes are easier to analyse than real imperfect planes, but the former still reveal something important about the latter. In normative theory, fictional cases of drowning children are (arguably) easier to analyse than complex real-world cases of global poverty with various empirical complications, but the former still reveal something important about the latter – because they reveal something about *one* particularly relevant normative factor at play in the real-world case.

Controlled Cases and Principle Testing

A principle is a statement that generalizes to more than one case.¹⁶ Because principles generalize, they enable philosophers to think about cases more systematically. Formulating principles naturally follows from decomposition: while the exercise of decomposing shows which factors might be relevant for the assessment of a case, well-formulated principles provide an account of how the different factors can be used to reach a normative or evaluative assessment.

One can think of a principle as a function, mapping each element of the *domain* (the set to which the principle is applicable) to one element of the *codomain* (the set of all possible assessments provided by the principle). Consider, for instance, a principle aiming to tell us which instances of defensive harming are permissible or even required. The domain may consist of all possible instances in which a person engages in defensive harm. The codomain consists of the three elements (impermissible, permissible, required). The principle, thought of

¹⁶ List, Christian, and Laura Valentini. “The Methodology of Political Theory.” In *The Oxford Handbook of Philosophical Methodology*, edited by Herman Cappelen, Tamar Szabó Gendler, and John Hawthorne, 525–550. Oxford: Oxford University Press, 2016.

as a function, determines for each possible instance whether this form of defensive harming is permissible, impermissible, or required.

A principle needs to pick up on patterns to be useful. To see this, think first of a maximally verbose and therefore not very useful principle: for each element in the domain it explicitly states which assessment from the codomain applies. This would result in a gigantic, potentially infinitely large lookup-table (“if this, then that...”) that provides an entry for every possible situation and the assessment of that situation. Needless to say, such a “principle” barely deserves the title. This is why the controlled cases described in the previous sections are so useful – if successful, they have already identified which properties can make a difference in the assessment, and which do not. The decomposed relevant factors allow the philosophical investigator to set aside most of the descriptive richness of the domain elements and instead focus on the small number of factors that make a difference. But most principles go further than that: instead of listing all possible combinations of factor instantiations, they give us a simple heuristic or formula, telling us which patterns of factors lead to which judgement.

In the seed case and decomposition stage, cases are used for exploratory purposes. But a key goal of moral theorizing is to formulate principles or even sets of principles that constitute theories. This leads us to two new functions of controlled cases: principle *testing* and principle *support*. We first turn to the role of principle *testing*, which is closely related to the question of case selection. Because a principle states a general relation between relevant factors and assessment, testing it requires that we choose cases systematically, mapping out the space of possible factor constellations. With unlimited time, we would want to map out the space systematically with a large sample of cases at many different locations. With more limited time, moral philosophers tend to select up to three different types of cases.

First, “corner cases” are situations in which one or more factor takes a (near) minimum or maximum value to test how the principle fares in the most extreme settings and assess its robustness. For an example of a corner case, take Nozick’s Utility Monster, which creates near infinite amounts of wellbeing from each unit of resources given to it.¹⁷ Utilitarianism seems to imply, therefore, that we should always give resources to the utility monster, rather than those who are much worse off, because this will maximise utility. Though unrealistic, the Utility Monster tests our judgements in a situation where the maximisation of utility is in extreme conflict with other possible values, such as equality or priority for the worse off. Corner cases give us an opportunity to test our commitments against extreme, even unrealistic pressure, in the same way plane wings are bent nearly 90 degrees in a stress test, even if they are unlikely to

¹⁷ Nozick, *Anarchy, State and Utopia*.

subject to such pressure during flight. Corner cases can also be counterexamples, the second type of controlled case often used for testing principles. Counterexamples often put moral judgments under pressure, but more generally they challenge principles by intuitions in the opposite direction. The Utility Monster is a corner cases, but it is also a counterexample because the intuitive judgement is that resources should go to the worse off rather than the monster, and thus the case suggests that Act-Utilitarianism is false. Third, controlled cases that can be used to observe the effect of one factor (or a small number of factors in interaction) while holding everything else constant, thereby attempting a strategy of isolation to observe singular effects (or factor interactions). We will return to some of these functions in the next section.

Argument Cases

We now begin to consider cases on the other side of the elicitation dimension: cases whose primary function is not the elicitation of intuitions but of reasons. We first turn to representative reason-generating cases, which we call “argument cases” and distinguish two different uses: for supporting principles and for testing metaphysical assumptions.

Argument Cases for Supporting Principles

Cases can lend support to principles in two ways: in *exposition*, by illustrating the application of the principles, or *substantively*, by demonstrating reasons that support the principles, though these two strategies of support often blend into each other. Cases of the former type are pedagogical devices for the benefit of the reader: stating the principle precisely would suffice to state the view, but an added example of application can support understanding, without necessarily supporting the content of the principle. Since this use of cases is familiar and fairly common, we won’t analyse it in more detail.

Cases that aim to provide substantive support for a principle go beyond mere illustration – they are also supposed to make the reader susceptible to the acceptance of reasons motivating the principle. For example, GA Cohen argues for his version of egalitarianism, and, more specifically, his interpretation of the difference principle, by providing an example. In his “kidnapper” case, Cohen asks us to imagine a criminal who has abducted a child and now tries to convince the parents to pay a ransom to him by insisting that children should be with their parents. Cohen points out that while this statement is generally true, the kidnapper is not in a

position to appeal to it as a premise of his argument. After all, the kidnapper is the cause why the child is not with their parents.¹⁸

The kidnapper case is interesting because it does not only elicit an intuition, it also makes the readers reason about the argument the kidnapper gives and why it fails. This demonstrates a new function of cases: apart from eliciting intuitions, some cases can also be used to elicit reasons to support an argument. When a case elicits reasons, it typically also elicits an intuition, but the intuition is not the goal of the exercise. In the kidnap case, for example, it is entirely unsurprising that we have the intuition that kidnapping is wrong, or that the reasoning provided by the kidnapper is preposterous. But the point of the kidnap case is to make the reader reason about the standing a speaker needs to have to make certain arguments. This insight is then used to criticize certain incentive-based arguments for demanding higher salaries.

Cases that elicit reasons will normally come with a richer logical structure than cases that elicit intuitions only. In Cohen's kidnapper case, the case itself contained an argument that provokes the reader into resisting the argument. Cohen also invites the reader to reason by structural analogy when comparing kidnapper with the case of a doctor who only works when they get a higher-than-average salary. Another common way to elicit reasons from cases is to compare two cases and analyse the difference between them.¹⁹

The distinction between cases for *testing* and for *supporting* principles allows us to state another principle of case use: testing and supporting cases must be chosen according to different criteria. Cases illustrate or support by eliciting reasons must be chosen for their ability to enable explanation, understanding and reasoning. They will be cases for which the application of the principle is most plausible, and they are chosen to make the assessment of the principle intuitive. The opposite holds for testing cases: they should be chosen to find out how robust the principle is in less paradigmatic case applications. That may involve exploring extreme assumptions or pro-actively scanning for counterexamples. Moreover, a meaningful test ought to be conducted by confrontation with several (and typically diverse) cases. Thus, the supporting and testing role should typically be fulfilled by different cases; running these two functions together would be a mistake.

¹⁸ Cohen, G. A. "Incentives, Inequality, and Community." In *Tanner Lectures on Human Value*, 1991.

¹⁹ Kimberley Brownlee and Zofia Stemplowska. "Thought Experiments." In: Adrian Blau, ed. *Methods in Analytical Political Theory*. Cambridge: Cambridge University Press, 2017.

Argument Cases for Querying Metaphysical Assumptions

The use of cases is not restricted to evaluative and normative investigations – it is equally important in conceptual analysis and even in metaphysics. Since ethical theory often depends on conceptual analysis or metaphysical assumptions, cases are often employed to test or query such assumptions. The use of cases for conceptual analysis have been analysed in detail elsewhere²⁰, so we set it aside in the interest of space. We do, however, investigate the use of cases for the analysis of metaphysical assumptions by looking at the metaphysics of causation and the metaphysics of harm. Cases of this type are often counter-intuitive: rather than being used to elicit intuitions, they show us that our intuitions and our background assumptions are in tension.

For an example of how ethics is influenced by the metaphysics of causation, consider overdetermination cases such as Parfit's two assassins:

“X and Y simultaneously shoot and kill me. Either shot, by itself, would have killed.”
(Parfit 1984, p. 70)

This raises questions about causation: whether X (or Y) has caused the death of hypothetical Derek. And entangled with this is the question if and why X or Y act wrongly, and whether X or Y are individually responsible for Derek's death. At the minimum, the case illustrates the questions to be discussed, but it also triggers judgements about both the causal and the ethical claims. The two assassins make us question common background assumptions about causation. For instance, a common understanding of causation is that the cause is necessary for effect. But that common understanding (together with some further auxiliary assumptions) leads to counterintuitive judgements about wrongfulness and responsibility in overdetermination cases, challenging the reader to revise either the background assumption about causation or the judgements about these cases.²¹

For an example of how the metaphysics of harm influences ethical theory, consider Warren S. Quinn's puzzle of the self-torturer.²² A patient can increase the electric current flowing through their body in tiny steps, such that the effect of each tiny increase is imperceptible, but come with a payment of \$10,000. The patient therefore prefers to nudge up the current at each step. However, once increased the current cannot be reduced, and once many steps have been taken,

²⁰ See, for instance, List and Valentini, “The Methodology of Political Theory”.

²¹ There is real disagreement about how these cases should be treated. See, for example, Jackson, Frank. “Which Effects?” In *Reading Parfit*, edited by Jonathan Dancy, 42–53. Oxford: Blackwell, 1997.

²² Quinn, Warren S. “The Puzzle of the Self-Torturer.” *Philosophical Studies* 59, no. 1 (1990): 79–90.

the pain becomes so unbearable that the patient would give up all his money to make it stop. This raises important questions about the analysis of harms that fall below the threshold of perceptibility. For instance, a common understanding of harm is that it must be directly perceptible. Another common understanding is that a relationship like “is as harmful as” is transitive, such that if A is as harmful as B and B is as harmful as C then A is as harmful as C. But these two assumptions (together with some further auxiliary assumptions) lead to the counterintuitive result that the lowest setting harms the self-torturer just as much as the highest setting, which is patently absurd. Either the assumptions or (less likely) the judgement must be revised.

For a case that tests underlying assumptions about identity and harm, take Parfit’s example-rich discussion of the non-identity problem. In Parfit’s description of *The 14 Year Old Girl*, the mother “would have had a different child” (Parfit 1984, p. 358), which is the starting point for a reasoning process about the fragility of identity and comparative notions of harming. The point of this case is, yet again, not only to elicit an intuition, but also to make the reader reason about the case. Parfit explicitly pursues both goals when discussing a related example: “My reaction is not merely an intuition. It is the judgement that I reach by reasoning” (Parfit 1984, p. 368), which is followed by a detailed engagement with the case to support Parfit’s view. This is a paradigmatic example for a case that forces us to analyse the background assumptions underpinning the relevant moral judgements and to potentially revise the judgements in light of what we learn.

What makes the cases for testing metaphysical assumptions so powerful is that they also have a representative role: our interest lies not in synchronized assassins, confused self-torturers, bean-stealing bandits, and perhaps not (or not exclusively) in the challenges of teenage pregnancy. Our interest arises because these construed cases represent larger classes of realistic cases that are highly relevant, and it is this power to represent that makes these cases relevant: they make us realize that some of the conventional thinking about applied, real-world cases might rest on muddled or at least questionable assumptions.

Cases for testing metaphysical assumptions typically play an auxiliary role in applied ethics and political philosophy by helping to investigate, clarify or revise background assumptions, though they can take centre stage in more theoretical projects. In the normal sequence of case use they are most useful after principles have been formulated. This is because they can serve as a check on the metaphysical assumptions made in the principle formulation. But in more theoretical projects, the case may be needed right at the start: to set up the puzzle and frame the debate. Which order works best depends on the context of the investigation and the division of labour

between theoretical and applied ethics. Interestingly, the debate about case use has largely overlooked this function of cases even though this category contains some of the most influential thought experiments appealed to in ethics.

Construction cases

Some of the most famous hypothetical cases in political theory play a role that we have not yet described. *Construction cases*, as we will call them, are used infrequently but often play a key role in grand theories. Two of the most famous construction cases are Rawls's original position and Ronald Dworkin's auction. Like argument cases, they seek to elicit the recognition of reasons, guiding the reader to understand, follow and accept arguments—albeit through a more complex modelling function. But unlike the cases in the last two categories, construction cases are specifically non-representative. They set out frameworks that constrain our reasoning and our judgements in particular ways, asking us to imagine a hypothetical, idealized choice situation—one that does decidedly not represent real-life choice situations—and to determine which outcomes would be accepted under such conditions.²³ The point of the construction case, then, is *not* to represent real choice situations, but to represent a plausible theoretical starting point that provides a focus for further normative theorising.

Construction cases can be understood as the final step, following the process of decomposing factors, organizing the factors into principles, and testing these principles against metaphysical and folk psychological assumptions. At this point, there will sometimes be factors, the strength of which a theorist is enormously confident about, but which people, in general, are likely to misjudge in their normative evaluations. Consider, for example, Rawls' original position. People are asked to imagine themselves behind a veil of ignorance that blinds them to their current position, privilege, and talents in society and decide upon principles for the societal distribution of benefits and burdens without such knowledge. The original position is “modelling the way in which the citizens in a well-ordered society, viewed as moral persons, would ideally select first principles of justice for their society”.²⁴ Rawls calls the original position a “device of representation”,²⁵ but he means a representation of these normative considerations. This is representation in a specifically normative sense—quite different from what philosophers of

²³ Bagnoli, Carla. 2011. “Constructivism in Metaethics.” Edited by Edward N. Zalta. *Stanford Encyclopedia of Philosophy*, doi:10.1111/1467-9973.00225.

²⁴ Rawls, John. 1980. “Kantian Constructivism in Moral Theory.” *The Journal of Philosophy* 77 (9): 520.

²⁵ Rawls, John. 1993. *Political Liberalism*. New York: Columbia University Press, 27.

science have in mind when they think about models.²⁶ When justifying the original position, Rawls states that it aims to ensure that “no one should be advantaged or disadvantaged by natural fortune or social circumstances in the choice of principles.”²⁷ The case accounts for these considerations, in other words, by incorporating into our reasoning a combination of factors, the normative significance of which Rawls is confident about—namely, equal concern for people’s claims regardless of background and abilities, or *fairness*.

The veil of ignorance makes vivid the underlying idea that the choice of principles should not be affected by arbitrary factors like unearned natural properties or pre-existing biases. Importantly, however, it also takes into account that people are likely to be affected by such factors and, thus, misjudge the fairness of potential principles of justice in ways that reflect their position and power in society. But as Rawls notes: “it should be impossible to tailor principles to the circumstances of one’s own case.”²⁸ The original position constrains our ability to do so. In principle, of course, we could appeal directly to fairness to argue in favour of Rawls’ principles. However, using fairness as a constraint on rational choice instead, inhibiting our ability to tailor principles to our own circumstances, captures the force of the argument in a different way—not least, by encouraging the reader to reach these conclusions from a first-person perspective.

Importantly, construction cases play a dual role in shaping our thinking by facilitating the strengthening of certain factors in our reasoning (e.g. fairness and opportunity costs) *and* helping to justify the principles and judgements reached via these cases by lending them added support. Thus, the hypothetical agreement itself constitutes an argument in favour of some principles (e.g. Rawls’ principles of justice) *because* the principles have been agreed upon in a choice situation that excludes partiality and ensures equal consideration of claims. Usually, discussions of construction cases focus solely on this principle-supporting output.²⁹ In DRE, however, we emphasize the double part that construction cases play in the process of justification. First, by using *input* from the previous stages to determine how our reasoning should be constrained. Second, by providing an additional, distinct underpinning for normative principles due to the controlled choice-situation into which the chooser is placed.

²⁶ Johnson, J. 2014. “Models Among the Political Theorists.” *American Journal of Political Science* 58 (3): 547–60, misses this important distinction in his discussion of models within political theory.

²⁷ Rawls, J. (1971). *A theory of justice*. Harvard university press, 18.

²⁸ Ibid.

²⁹ E.g. Brownlee & Stemplowska (2017); Brun (2017); and Knight, C. (2017). Reflective equilibrium. *Methods in Analytical Political Theory*, 46–64.

Defence of Directed Reflective Equilibrium

Seed cases can be hypothetical but are often better taken from real life, for two reasons. First, one criticism of hypothetical cases is that they problematically abstract from real world issues. We will address this criticism when we talk about decomposition and how this underpins the representative function of cases, but for now it is important to note that, even if cases can successfully represent similar factors in different contexts, seed cases provide the initial impetus for determining what those factors are. They establish the boundaries of the moral phenomenon under investigation. Beginning the enquiry with a real-world seed case, however complex or “murky” from an analytic perspective, helps to focus the enquiry on the salient moral factors – or, as Susanne Burri puts it in a recent article, starting from real-world seed cases helps ensure “practical applicability”.³⁰ This enables the directional approach to address one of the problems we previously noted with regard to RE. RE does not prescribe any specific starting point for moral theory. A theorist might start from a specific case but might equally start from an abstract principle. The use of seed cases in DRE, by contrast, represents an attractive middle ground between fixating on specific real-world problems and pursuing highly abstract theory.

Second, such seed cases have a better chance of maintaining real-world representativeness. If one begins from a real case, after the following steps are completed (see sections XXXXX), there is a higher likelihood that resulting principles will maintain their representative connection to the real moral phenomenon. This also helps DRE address the problem of abstractionism. Further variations on cases, which tend increasingly towards the creation of far fetched or fantastical examples, are more likely to maintain their representativeness if they are based on seed cases rather than beginning immediately with cases designed to isolate one moral factor from all others.

³⁰ Burri, S. (2019). Why Moral Theorizing Needs Real Cases: The Redirection of V-Weapons during the Second World War. *Journal of Political Philosophy*.

It is important to keep in mind, furthermore, that the point of this systematic analysis is not just to pull apart complex cases, but to stitch them back together again and look at them anew, hopefully with a deeper understanding of their moral complexity. To evaluate the case of vaccination campaigns, for example, we might also want to explore factors like the role obligations of doctors and nurses, how to ensure continued trust in the health system, and upholding political ideals of equal respect and concern for individuals.

This is significant because, although seed cases have intuitive pull, the intuitions they elicit are frequently muddled and obscured by being bundled up in complex ways. Multiple normative factors often coexist, making it difficult to appreciate which judgments or reasons, if any, are supported by which factor. Because of this, it is often valuable to analyse cases in which moral considerations that typically coexist are separated to see how they function independently. This often requires constructing controlled cases since in most realistic cases the considerations that we wish to pull apart are found together. Controlled cases offer a useful analytic tool to achieve this. When faced with a complex, perhaps real world, moral case, we are presented with a choice: we can either evaluate the case in all its complexity, attempting to discuss relevant considerations without comparison with other cases. Alternatively, we can tease apart different factors by considering other cases where these factors are present, but others that co-existed with it in the original case are absent. Thus, a single complex case can become a family tree of cases.

Controlled cases like Drowning Child or Trolley thus deliberately aim to test or support the importance of specific factors by isolating their intuitive pull and suppressing the effect of other factors. Factors are explored, then, by eliciting intuitions about them individually (or, if necessary, in deliberate interaction with other factors) and good hypothetical cases are ones that both represent factors present in a number of real life cases *and* elicit clear intuitive responses. As mentioned, Drowning Child is inspired by an actual famine in South Asia. Alleviating actual famines by donating money to charities, of course, does not happen as straightforwardly as does saving the child in Singer's example. Many have raised worries about factors which are relevant when considering charitable donations that are not present in Drowning Child. Some worry,

for example, that, unlike saving the drowning child, charitable donations are often ineffective, create and uphold relations of dependency, help sustain corrupt governments, and that they do not suffice to remedy global poverty and injustice.³¹

In the role controlled cases are meant to play in DRE, however, Drowning Child is not *meant* to include these factors because it is not meant to replicate the normative complexity of an actual famine. Rather, it is meant to isolate and foreground the intuitive pull of one factor—being able to help others greatly at little cost to oneself. In this particular example, the case is also meant to suppress another factor, which is present and which is often given exaggerated importance in cases of actual charitable donations—geographical distance. Drowning Child does not tell us what to do when faced with an actual famine, but it helps us untangle the complexity of the situation by highlighting factors that we are liable to underappreciate and subduing other factors, the importance of which we are liable to overestimate (such as geographical distance). More generally, then, hypothetical cases representing decomposed factors can help provide clarity about the real-life dilemmas of seed cases, in which factors are intertwined and obscured.

Controlled cases will sometimes require an unrealistic setup in order to isolate the relevant factors.³² Consider Thomson's 'people-seeds' example.³³ In this case, people-seeds drift about in the air like pollen, and despite the mesh screens erected to prevent their entry, they take root in the carpet and start to grow, eventually turning into human beings. Though this example is absurd, it is intended to be analogous to pregnancy via intercourse that one has taken reasonable steps to avoid. Since there are no realistic cases of this kind (except actual pregnancies that cannot function as analogies) the analogy is necessarily fantastical. Again, the case does not give us conclusive evidence about the real-life dilemma from which the factor is drawn—the permissibility of abortion. It does, however, provide information about *one* important factor of such dilemmas: the extent to which we can incur demanding, individual obligations to sustain potential human life when we have taken all reasonable steps to avoid this potentiality.

Compared to standard reflective equilibrium, our approach has two advantages. Both stem from the fact that DRE distinguishes clearly between seed cases (not to be confused with Thomson's people-seed case), drawn from real life, and hypothetical controlled cases and how and when the two are best used. First, in standard reflective equilibrium, real and hypothetical cases are

³¹ Miller, D. (2007). *National responsibility and global justice*. Oxford University Press, chapter 9; Unger, P. K. (1996). *Living high and letting die: Our illusion of innocence*. Oxford University Press, USA.

³² Brownlee, K., & Stemplowska, Z. (2017). Thought Experiments. *Methods in analytical political theory*, 21-45.

³³ Thomson, J. J. 'A Defence of Abortion', *Philosophy & Public Affairs*, Vol. 1, no. 1 (Fall 1971).

not clearly distinguished as sources of intuitions or considered judgements, but play the same role in the process of building and testing principles. On our view, while both types of cases elicit intuitions, the two play different roles in the justificatory process. Intuitions drawn from real-life cases will often less reliably track particular factors because such cases are messier and, thus, it will be less clear which factor(s) is causing the elicitation.³⁴ Controlled cases, on the other hand, can isolate specific factors and elicit targeted intuitions to test their force. Our intuitions about real-life cases, then, are made clearer when the various factors are decomposed and reconstructed into factor-representative controlled cases.

Second, standard reflective equilibrium does not include the precept of beginning or grounding one's normative exploration in seed cases. On the standard approach, intuitions drawn from hypothetical examples can, in principle, be entirely unconnected to real-life situations. This gives rise to the worry that such intuitions have little bearing on actual moral and political dilemmas. Thus, while intuitions elicited by hypothetical cases better track individual normative factors, the guidance such intuitions provide for moral and political agency is limited, if the hypotheticals are not grounded in real life. Our method seeks to alleviate this shortcoming by integrating a basis in reality by urging theorists to start from seed cases and keep a constant eye on how different factors under consideration appear in actual dilemmas (sometimes, this will happen by urging theorists to work backwards from a seemingly relevant factor to seed cases in which this factor exists).

Working with Controlled Cases, philosophers aim at discovering regularities and to capture these regularities in *principles*. We will now turn to the role of principles in reflective equilibrium and how principles and cases interact.

Unsurprisingly, thought experiments come into their own when such test cases are required. Thought experiments can be used to consider situations that can have never arisen in real-life (or are unknown to the philosopher). They can help construct “clean” controlled cases that focus only on the factors under consideration. And they easily lend themselves for thinking

³⁴ Tadros, V. (2011). *The ends of harm: The moral foundations of criminal law*. OUP Oxford, pp. 7-10.

Commented [ADV1]: Insert better transition to principles section

Commented [KS2R1]: Like the two sentences below?

Commented [ADV3]: This section could be shortened

about extreme or “corner” cases, or for constructing a series of cases to be considered in comparison such that only one factor varies.

The idea of controlling for factors can be observed in the large class of *Trolley* cases and their comparison. For instance, several variations of Trolley explore whether harm as a side effect and harm as a means to an end ought to be assessed differently, thereby testing the Doctrine of Double Effect. Counterexamples are also a popular choice of case because a convincing counterexample provides strong evidence against a principle. Nozick’s Utility Monster, for instance, can also be read as a counterexample to some forms of **utilitarianism**.

~~Despite all these advantages of thought experiments, real-world cases also have their place in principle testing. Going back to the initial seed cases helps philosophers to assess whether the principle works for the richer descriptions the seed cases provide. And in some instances, revisions in the initial assessment of the seed cases are required to achieve equilibrium.~~

~~[a case study of a mistake would be nice here! Perhaps an example from the history of philosophy, not necessarily in ethics? Or even an analogue from the sciences?]~~

Commented [KS4]: We could delete this paragraph as it revisits controlled cases. Maybe use further up?

Commented [KS5]: CUTOUT: Despite all these advantages of thought experiments, real-world cases also have their place in principle testing. Going back to the initial seed cases helps philosophers to assess whether the principle works for the richer descriptions the seed cases provide. And in some instances, revisions in the initial assessment of the seed cases are required to achieve equilibrium.