Directed Reflective Equilibrium: Thought Experiments and How to Use Them

PRE-PROOF VERSION

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 wide 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 to test principles and elicit moral judgements. 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. There are numerous criticisms, but here are two familiar ones. First, 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.[3] A suspicion of abstractionism underpins much historical scepticism towards moral theory in general,[4] and similar worries can be raised about hypothetical cases. Second, it is often said that RE relies on a coherentist approach to justification: the idea that the coherence of a set of beliefs justifies these beliefs. But, if the RE is read as an instance of coherentist justification, it faces a challenge about what to do in the event of inconsistency between our intuitions or between intuitions and basic principles.[5] To be sure, there are theoretical resources to overcome this impasse: the robustness of judgments, the vulnerability of intuitions to debunking, theoretical parsimony, and so on. However, what RE is still lacking is a sense of how cases ought to be *sequenced* in theoretical enquiry, given their different uses. Distinguishing between different types of cases and

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

developing a sensible model for sequencing them within theoretical enquiry helps to avoid some of the pitfalls of the case-based methodology, or so we will argue.

We aim to 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 overcoming some of the methodological shortcomings faced by RE. With a clearer typology of cases in mind a sequence of their usage suggests itself, which helps overcome the pitfalls of RE. In referring to a 'sequence', we mean using different cases for different purposes at different stages of a theoretical enquiry – engaging in directed rather that non-directed RE. We do not suggest the use of cases should be rigidly sequenced: some stages may be omitted, and DRE accommodates a degree of movement back and forth between different stages of analysis in the manner of RE. Nevertheless, we will argue that DRE has a number of advantages over a non-directional approach to RE.

The suggested sequence of 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 "decomposed" into various moral factors that might affect our intuitions. Here, we understand moral factors as facts that have some weight or relevance in considering what an agent ought to do. Decomposition allows the philosopher to construct "controlled cases" that represent moral factors, independent of both the original context of the seed case and the other factors with which it previously coexisted. Testing different versions of these cases against each other, the philosopher then seeks to "organize" 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 as well as intuitions, seeking to support principles on the one hand, and challenge biases, metaphysical beliefs, and underlying conceptual assumptions that may colour our intuitions on the other. Fourth, principles that cohere with both intuitions and reasons can be "veiled" in the final type of case. "Construction cases" set up choice situations incorporating fundamental principles, making choices that do not accord with these principles impossible.

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 two sections we explain our taxonomy and develop the model in greater detail. We then argue that the model improves on RE by addressing some of the pitfalls of the case-based methodology mentioned above.

A Two-Dimensional Case Typology

An important distinction for understanding the case typology we introduce in the next section is between two dimensions of cases: representation and elicitation. We now explain these dimensions before showing how they structure the process of DRE.

The Representation Dimension

We explain the representation dimension of cases by borrowing from the discussion of models in the philosophy of science.[6] 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. 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, of course, an idealized exemplar of real cells,[7] but what makes it similar is that certain structural features are alike.[8]

The sciences use models in the form of equations, computer code or scale models. In normative theory, however, most models are "word models," stated purely in narrative form.[9] 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, for example, frictionless planes are easier to analyse than real imperfect

---

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

[7] 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.

[8] One popular approach in the sciences has a "hub-and-spoke" structure, as Peter Godfrey-Smith, 'Models and Fictions in Science', p. 107, 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."

[9] The occasional formalized game-theoretical model can be found but remains the exception rather than the norm.

planes, but the former still reveal something important about the latter. In normative theory, fictional cases are (arguably) easier to analyse than complex real-world problem. Take Peter Singer's famous example, Drowning Child,[10] in which is it possible to saving the life of a child–but only by getting an expensive pair of shoes wet. This case helps us approach issues of global poverty. Of course, such issues are much more complex and have various empirical complications. But the former reveal something important about the latter—because they reveal something about *one* relevant normative factor at play in the real-world case (we will discuss concerns with the representativeness of cases such as Drowning Child later).

Cases used as models are really idealized exemplars: models of a larger class of real-world cases. Just like the drawing of the cell, the models are unrealistic due to high idealization. But they are unrealistic for a purpose: to single out structural aspects that they would share with all the real-world cases they represent. This, then, sets up the real-world representation dimension: cases are either representative or non-representative of normative factors that we must incorporate into our deliberation when facing real world situations. As we will see when we introduce our sequence, the representativeness of cases informs how they should be constructed and selected.

The Elicitation Dimension

Cases can be used for different purposes. In particular, they can be used to trigger different responses. On the one hand, cases can be used to elicit intuitions. On the other hand, there are cases that are not, or not primarily, used to elicit intuitions but rather to elicit the recognition of reasons. Sometimes the case offers an argument to the audience and tries to convince them of its correctness. Then the response is the acceptance (or rejection) of the argument. Sometimes the case is constructed to encourage the audience to reason towards an argument. Either way, after the case has been presented, the audience is supposed to relate to reasons, not just report an intuition.

Consequently, on the elicitation dimension, a case can either trigger intuitions or reasons. Having specified these two central dimensions upon which cases used in normative theorizing differ, we can now flesh out how these dimensions figure in the process of DRE.

A Typology of Cases

Figure 1 shows the case taxonomy we are proposing, with the two dimensions indicated at the top and left of the figure setting up a 2x2 typology. The four types of cases we distinguish all have a role to play in DRE. In the next section we explain these case types and their function

[10] Singer, Peter. 'Famine, affluence, and morality.' *Philosophy & public affairs* (1972): 229-243.

in terms of representation and elicitation. The ideal sequence of DRE is indicated by grey arrows, showing that, in the most complete DRE process, one starts with seed cases, proceeds to controlled cases and argument cases, and ends with construction cases. Finally, figure 1 also points to the central role of principles. All but the seed cases have a function related to the formulation, testing, support and systematization of principles.

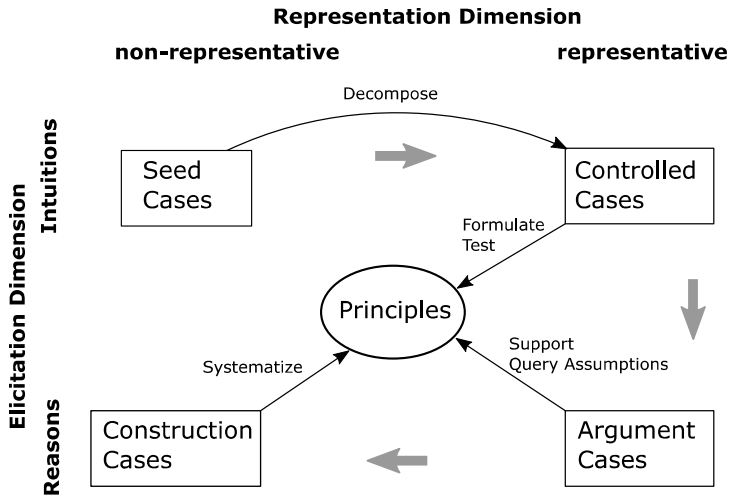

Figure 1: Directed Reflective Equilibrium Case Use.

Seed Cases and Decomposition

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,[11] cases of famine or other humanitarian crises may inspire discussion of the duty of rescue,[12] 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

[11] For example, military acts designed to terrorise a population into surrender provide the foundation for the comparison between 'terror bomber' and 'tactical bomber' cases. See, for example, Walzer, M. (1971), 'World War II: Why Was This War Different?', Philosophy & Public Affairs 1/1: 3-21.

[12] Singer, 'Famine, affluence, and morality.'; Gerver, Mollie. *The ethics and practice of refugee repatriation*. Edinburgh University Press, 2018.

complex moral problems involving, among other things, the distinction between positive and negative duties, the stringency of assistive duties, the 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 multi-faceted moral problems. They pull our intuitions in different directions, perhaps in accordance with pre-existing moral or political sensibilities, and almost always involve a complex intersection of different morally salient facts. In our case typology, seed cases are intuition-eliciting and non-representative: they provide the basis for constructing other, simpler cases, which either represent elements of the seed case or elicit normative responses to factors drawn from the seed case.

The complexity or "murkiness" of seed cases can be daunting. 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 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 be separated from the others and formulated into further cases. Many revisionists in just war theory, for example, compare 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.[13]

Controlled Cases

Building on the output of the decomposition, philosophers can systematically integrate the different factors into hypothetical 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

[13] For a key text in revisionist just war theory that takes this approach, see McMahan, Jeff. *Killing in War* (Oxford: Oxford University Press, 2009).

to cases used in this stage of the process of DRE as *controlled cases* to emphasise their use in separating factors.[14] Unlike seed cases, which are singular, however, these cases aim to *represent* a particular factor that is present in many real life situations. To demonstrate how controlled cases work, consider the famous trolley case (Trolley). In this case, we must decide whether to do nothing and allow a runaway trolley to kill five people, or to divert the trolley onto a sidetrack where it will kill one person. Trolley is an interesting moral dilemma in itself, but the feature we highlight here is that Trolley aims to represent a factor that is present in a wider class of cases. When debating Trolley and its variations, we're clearly not interested in railways, trolleys, people tied to tracks, and the like. Rather, we are interested in a large class of cases in which killing or letting some die can save a larger group of others. This normative factor is what Trolley seeks to bring to the fore, and the numerous variations of Trolley do the same with other factors. In other words, trolley cases serve as a stand-in for many real-world cases with similar structures and the factor emphasized in Trolley is representative of a larger class of cases that are of genuine real-world interest.

Testing and Supporting Principles

Controlled cases can then be used to test principles. A principle is a statement that generalizes to more than one case.[15] Because principles generalize, they enable philosophers to think about cases more systematically. Formulating principles naturally follows from decomposition: while the exercise of decomposition 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

[14] 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.

[15] 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. Our understanding of principles is consistent with treating principles as summaries of normative facts rather than grounds of normative facts. See Berker, S. (2019), The Explanatory Ambitions of Moral Principles. Noûs, 53: 904-936.

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 philosopher 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 stages, cases are used for exploratory purposes. But a key goal of moral theorizing is to formulate principles or sets of principles that constitute theories. This leads us to two new functions of controlled cases: principle *testing* and principle *support*. We first address the role of principle *testing*, which is closely related to the question of case selection, then turn to principle support in the next section. 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 factors take 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.[16] 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 worst off. Corner cases give us an opportunity to test our commitments against extreme, even unrealistic pressure, in

[16] Nozick, Robert. *Anarchy, State and Utopia* (New York: Basic Books, 1974).

the same way plane wings are bent nearly 90 degrees in a stress test, even if they are unlikely to be subject to such pressure during flight. For similar reasons, we should be interested in the *robustness* of a principle's plausibility, rather than just its plausibility in the range of cases we are most likely to confront.

Corner cases can also be counterexamples, the second type of controlled case often used for testing principles. Counterexamples put moral judgments under pressure, but more generally they challenge principles by intuitions in the opposite direction. The Utility Monster is a corner case, but it is also a counterexample because the intuitive judgement is that resources should go to the worst off rather than the monster, and thus the case suggests that Act-Utilitarianism is false. Finally, controlled cases 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 when we outline the benefits of DRE.

Argument Cases

Supporting Principles

Argument cases are not employed in a purely exploratory mode – they also have an argumentative function. We draw attention to two important argumentative purposes: for supporting principles and for testing metaphysical assumptions. Take supporting principles first. 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 example of its application can support understanding, without necessarily supporting the content of the principle. For example, Trolley may be used to illustrate the difference between Act-Utilitarianism and the Principle of Doing Allowing by pointing to their different implications with respect to permissibly diverting the trolley.

Cases that aim to provide substantive support for a principle go beyond mere illustration – they are also supposed to incline the reader to accept the principle. Of course, there is no rigid distinction between controlled cases and argument cases, between exploration and argument. Two types of cases mentioned above – corner cases and counterexamples – have an important role to play in arguing against or in favour of principles. But argument cases have other functions, too. 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 of the child not being with their parents.[17]

The kidnapper case is interesting because it does not only elicit an intuition, it also encourages the reader to 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 the recognition of reasons to support an argument, marking the next dimension shift in our typology. When a case elicits the recognition of a pattern of reasoning, it typically also elicits an intuition, but the intuition is not necessarily the goal of the exercise. In the kidnapper 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 kidnapper case is to make the reader reason about the standing a speaker needs to have to make certain arguments. This insight is then transposed to a different context and 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 the kidnapper with the case of a doctor who only works when they get a higher-than-average salary: a common way to elicit reasons from cases is to compare two cases and analyse the difference between them.[18]

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 that illustrate, or support by eliciting reasons, should 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

[17] Cohen, G. A. "Incentives, Inequality, and Community." In *Tanner Lectures on Human Value*, 1991.
[18] Kimberley Brownlee and Zofia Stemplowska. "Thought Experiments." In: Adrian Blau, ed. *Methods in Analytical Political Theory*. Cambridge: Cambridge University Press, 2017.

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.

Querying Metaphysical Assumptions

The use of cases is not restricted to evaluative and normative investigations – it is equally important in conceptual analysis and 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 elsewhere[19], so we set it aside in the interest of space. We will, however, briefly demonstrate the use of cases for the analysis of metaphysical assumptions by looking at the metaphysics of causation and the metaphysical assumptions related to different conceptions 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 Derek 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. And entangled with this is the question whether 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 assumption about causation is that the cause is necessary for the effect. But that assumption (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.

For an example of how the metaphysical assumptions concerning harm influence ethical theory, consider Warren S. Quinn's puzzle of the self-torturer.[20] 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 comes with a payment of $10,000. The patient therefore prefers to nudge

[19] See, for instance, List and Valentini, "The Methodology of Political Theory".
[20] Quinn, Warren S. "The Puzzle of the Self-Torturer." *Philosophical Studies* 59, no. 1 (1990): 79–90.

up the current at each step. However, once increased the current cannot be reduced, and once many steps have been taken, 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 assumption about harm is that it must be directly perceptible. Another common assumption 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 absurd. Either the assumptions or (less likely) the judgement must be revised.[21]

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, and so on. Our interest arises because these cases represent larger classes of realistic cases 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 normative theory play a role that we have not yet described. *Construction cases*, as we will call them, are used infrequently but often play a

[21] Other examples of argument cases for querying metaphysical assumptions include the bean-stealing bandits in Glover, J, and M J Scott-Taggart, "It Makes No Difference Whether or Not I Do It." *Proceedings of the Aristotelian Society*, Supplementary Volumes 49 (1975): 171–209 and the pregnant 14-year old girl in Parfit, Derek. *Reasons and persons*. OUP Oxford, 1984, §122.

key role in grand theories. One of the most famous construction cases is Rawls's original position. 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 decidedly does not represent real-life choice situations—and to determine which outcomes would be accepted under such conditions.[22] 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. They fill the last remaining cell of our typology.

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 salience of which a theorist is very confident about, but which people are generally 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".[23] Rawls calls the original position a "device of representation",[24] but he means a representation of these normative considerations. This is representation in a specifically normative sense—quite different from what philosophers of science have in mind when they think about models.[25] 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."[26] 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*.

---

[22] Bagnoli, Carla. 2011. "Constructivism in Metaethics." Edited by Edward N. Zalta. *Stanford Encyclopedia of Philosophy*, doi:10.1111/1467-9973.00225.
[23] Rawls, John. 1980. "Kantian Constructivism in Moral Theory." *The Journal of Philosophy* 77 (9): 520.
[24] Rawls, John. 1993. *Political Liberalism*. New York: Columbia University Press, 27.
[25] 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.
[26] Rawls, J. (1971). *A theory of justice*. Harvard university press, 18.

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."[27] 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.

Other examples of construction cases include: Ronald Dworkin's auction, in which we are asked to imagine a group of shipwreck survivors washed up on an island and faced with the task of dividing the island's land and resources in a just manner among themselves through an auction which is meant to leave everyone content with their post-auction bundle;[28] Adam Smith's impartial spectator, which asks to evaluate the moral sentiments of ourselves and others from the point of view of a well-informed and impartial spectator;[29] Dworkin's judge Hercules, which asks how an ideal judge with unlimited time and knowledge would rule on constitutional cases; [30] Chandran Kukathas' 'liberal archipelago', which considers society as a collection of co-existing but separate societies and asks which rules should guide such diversity;[31] and Thomas Hobbes' state of nature, which highlights the dangers of living without (and the difficulties of achieving) stability. All of the mentioned cases function by using factors as input to constrain our reasoning, ensuring, for example, that we take into account the dangers of societal instability or the opportunity costs of our choices, or that we disregard partiality towards our own situation. By incorporating factors in this way, construction cases *exclude* certain normative conclusions from being reached. This can help explain that such cases are used so rarely—because excluding certain conclusions requires an extraordinarily high level of confidence in the relevant, excluding factor.

---

[27] Ibid.
[28] Dworkin, R. (1981). What is equality? Part 2: Equality of resources. *Philosophy & public affairs*, 283-345.
[29] Fleischacker, S. (2013), Adam Smith's moral and political philosophy, *Stanford Encyclopedia of Philosophy*. This notion is also used in the construction and justification of Roger Crisp's account of sufficiency. See Crisp, R. (2003). Equality, priority, and compassion. *Ethics*, 113(4), 745-763.
[30] Dworkin, R. (1986). *Law's empire*. Harvard University Press.
[31] Kukathas, C. (2003). *The liberal archipelago: A theory of diversity and freedom*. Oxford University Press.

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.[32] In DRE, however, we emphasize the double role 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.

Some Advantages of Directed Reflective Equilibrium

In this section we outline some of the specific advantages of the DRE over non-sequenced approaches. The first advantage was made evident, we hope, in the course of laying out the model. Cases have a multitude of purposes in normative theory, and though these are often recognised implicitly, it pays to have a more explicit and comprehensive taxonomy. This allows us to construct cases in accordance with their specific purpose, and, as previously noted, this may affect the content of the case because the criteria for case selection vary depending on the type of case being constructed.

DRE is not merely a taxonomy of cases, however: it also recommends a specific progression of case-use in normative theory. Having outlined the various functions of cases, we can see how, overall, the use of cases moves from an exploratory mode to an argumentative mode. Many cases in philosophical writing have an argumentative purpose: they aim to pump intuitions, provide counterexamples, and so on. By contrast, DRE encourages us to make use of cases just as much in our early exploratory phase as in our later argumentative phase. And having distinguished between the various functions of cases, we are now better placed to see how the use of cases can go wrong when the two phases are mixed.

When philosophers move to the testing of principles too quickly, this narrows the inquiry in two ways. First, skipping the exploratory seed case and decomposition phase increases the risk of missing important factors and fixating too quickly on existing principles. Reducing the list of

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

candidate factors for principle formulation narrows the scope of the search for new principles, especially those principles that are not easily identified due to bias or inertia. This danger is particularly relevant in applied ethics, and especially when investigating new or philosophically under-explored phenomena. But even in well-established areas of philosophical research it is important not to rule out overlooked factors too early. One may be tempted to select a factor that leaps out at us from a seed case and consider this in more detail. Philosophical debate often operates in this way, where one thinker will highlight a factor that appears important and another will criticise this and highlight a different factor. Decomposition encourages us to begin by simulating this dialectical process intrapersonally before defending any one factor, by breaking down the seed case into as many relevant factors as possible.

It is worth noting that we are describing a process of philosophical *thinking* rather than *writing*, and some or even all of the early stages may not be incorporated into written output. However, there are a number of benefits of making this process explicit and, in particular, of performing it early in the reflective equilibrium process. First, decomposition is a more neutral way to capture the variety of factors with potential moral relevance. A common criticism of appeal to moral intuition is that our intuitions are shaped by biases and pre-existing theoretical commitments. Decomposition offers a way to mitigate this worry by extracting as many moral factors as possible from a seed case and formulating them into cases of their own. Of course, no methodology can eliminate bias entirely, but engaging in thorough decomposition before using argument cases is one way to guard against selection bias.

Second, if relevant factors are not identified in the exploratory phase, the preconceived principles are unlikely to be tested with cases that present variations on these factors, either separately or in interaction. Moving on to testing principles before a careful exploratory phase has the counter-productive effect that the testing will be less comprehensive because alternative hypotheses are not explored. The appropriate use of argument cases to test metaphysical and conceptual assumptions also helps the directional approach to avoid path-dependency problems. The conclusions arising from normative theory depend, in part, on the metaphysical assumptions we adopt. These cases ensure that the results of our theorising are more likely to determine whether intuitive disagreement is based on genuine normative disagreement or disagreement about metaphysical or conceptual assumptions.

An Anchor in the Real World

One common criticism of hypothetical cases, which we sketched at the outset, is that they are abstract or fantastical and therefore not relevant to real world problems. There are plenty of

responses available to this charge;[33] here we add one more. DRE recommends beginning enquiry with a seed case. Such cases, however complex or "murky" from an analytic perspective, help us focus on the salient moral factors that we find in real world scenarios and therefore "anchor" the ensuing enquiry. Or, as Susanne Burri puts it in a recent article, starting from real-world seed cases helps ensure "practical applicability".[34] This can help to ensure that the results of philosophical enquiry have implications for what we (as individuals, groups or states) ought to do with regard to these problems, as long as subsequent stages of theorising are also performed with care, e.g. ensuring controlled cases maintain their representativeness with the seed case, even if they transpose a factor into a very different context.

Normative theory that begins with discussion of abstract principles may still have practical implications: utilitarianism, for example, has many practical implications. But practical implications are not the same as practical applicability, or, in our terms, anchorage in real world problems. The use of seed cases helps to focus our attention on moral phenomenon that are pertinent to real-world moral issues, beyond ensuring that the results of theorising have practical implications. 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.[35] 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.[36]

On the standard approach to reflective equilibrium, 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. If one begins from a seed case, after the following steps are completed,

---

[33] See, for example, Brownlee and Stemplowska (2017). For a moderate defence of thought experiments, see Walsh, A. (2011). A Moderate Defence of the Use of Thought Experiments in Applied Ethics. Ethical Theory and Moral Practice, 14(4), 467-481

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

[35] As Eva Erman and Niklas Möller put it in "Practices and principles: On the methodological turn in political theory." *Philosophy Compass* 10, no. 8 (2015): 533-546, reflective equilibrium "is completely neutral with regard to where we start our normative endeavour – we may start with abstract principles or with local norms and contextual claims."

[36] Very recently, Eric Brandstedt and Johan Brännmark have suggested a way of making reflective equilibrium more practical by combining it with Rawlsian Constructivism in "Rawlsian Constructivism: A Practical Guide to Reflective Equilibrium." *Journal of Ethics* (2020). Our approach adds more specification to the role played by cases and the sequencing of different stages of the method, but otherwise we take the two approaches to be compatible.

there is a higher likelihood that resulting principles will maintain their representative connection to the real moral phenomenon.

Distilling Clarity from Complexity

Although seed cases may 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 unrealistic cases since in most realistic scenarios the considerations that we wish to pull apart are found together. Decomposition and controlled cases are useful tools for achieving this. Including decomposition as an explicit stage of the enquiry models something that often emerges dialectically: an itemisation of the various moral factors that may play a role in the seed case. Controlled cases then offer a useful analytic tool to separate factors from their original context to see how they operate independently. 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. 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.[37]

[37] 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.

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.

It is important that controlled cases properly represent the factors they draw from the seed case. However, we must also be clear that controlled cases represent a *factor*, not the seed case itself. As mentioned, Drowning Child represents the ability to save life at low cost, and thus excludes other (perhaps important) factors from real famines such as geographical distance. According to DRE, it is irrelevant for critics to focus on the various ways a controlled case is unlike the real phenomenon in which the theorist is interested. It is far more important to conduct the process of decomposition thoroughly, ensuring that factors are properly articulated, to maintain representativeness with the seed case.

Once we understand the rationale behind decomposition and controlled cases, we can be clearer about the criteria for case construction. We should begin by considering how best to separate a factor from a seed case with minimal distraction. No more contextual information should be added to the controlled case than is necessary to maintain representativeness with the relevant factor from the seed case. We should then ask whether the benefit of representing this factor outweighs the distraction.

Controlled cases will sometimes require an unrealistic setup in order to isolate the relevant factors,[38] but this should always be balanced against the benefit of representation and isolation. Consider Thomson's 'people-seeds' example in this regard.[39] 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

[38] Brownlee, K., & Stemplowska, Z. (2017). Thought Experiments. *Methods in analytical political theory*, 21-45.
[39] Thomson, J. J. 'A Defence of Abortion', *Philosophy & Public Affairs*, Vol. 1, no. 1 (Fall 1971).

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 issue 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. Our discussion shows how people seeds, despite its fantastical nature, does precisely what controlled cases ought to do. It isolates a particular variable and puts it into a context that can properly function as an analogy. The fantastical elements are excused since it is difficult to see how cases that capture the relevant factors could be more realistic without involving actual pregnancies.[40] Perhaps there are cases that successfully represent the same factors whilst being less mired in fantastical detail. We do not know what such cases would look like, but our present purpose is to articulate the criteria that should govern the construction of such cases: maintaining representativeness with the factor taken from the seed case with minimal possible distraction through fantastical detail.

Spotting Interaction Effects

Decomposition is effective for identifying interaction effects. Frances Kamm sums up the general phenomenon of interaction effects with her Principle of Contextual Interaction. According to this principle, a factor's moral salience may differ with its context. If a factor seems irrelevant in one case, it doesn't follow that it is irrelevant in others, and vice versa.[41] One worry about RE is that it has no inbuilt mechanism to detect interaction effects. A principle may be consistent with one intuition about a specific factor and thus be in RE, but the Principle of Contextual Interaction suggests that equilibrium may yet be threatened if the intuitive judgment about that factor changes when it is transposed to a different context.

To demonstrate, consider again the case of killing a combatant in war. We might reason that culpably causing a threat to another is a ground for liability to defensive harm. But it would be too hasty to conclude from this anything about the relevance of causing a threat and being culpable independently. It may be that causation on its own is relevant, or perhaps not; the same is true for culpability. Until we have separated these factors and formulated them into their own cases systematically, we run a greater risk of drawing conclusions without taking into account interaction effects. Using decomposition to separate factors from seed cases, and therefore from their original context, helps to isolate those factors. And once this isolation is achieved, it is

[40] For a similar argument, see REDACTED.
[41] See Francis Kamm, *Morality and Mortality Vol. 2.*, (Oxford: Oxford University Press, 1996), 51.

possible to conceive of controlled cases that deliberately vary a select number of factors to pick up interaction effects early.

Different Forms of Elicitation

Reading many philosophical debates, one might be forgiven for thinking that the primary argumentative purpose of cases is to pump intuitions.[42] This is, indeed, an important function, but DRE employs cases for various forms of elicitation beyond intuition pumping. Cases can either trigger intuitive responses, or they can be used for argumentative purposes, perhaps as presumptive support for a philosophical claim or to explore the implications of different principles. Such reason-eliciting cases can also be employed to query assumptions and refine principles. These different purposes are part of the basis for distinguishing between various argument cases. And as our discussion of Rawls' original position shows, construction cases also elicit patterns of reasoning.

We hope our account of the DRE model highlights the benefit of these different forms of elicitation. Of course, RE might also make use of these cases, and in this respect, they are not exclusive to a directional approach. Nevertheless, we suggest that they are best placed after the exploratory process of decomposition, and the investigation of moral factors in independent controlled cases. The different forms of elicitation reflect cases in their argumentative mode: used for pumping intuitions; encouraging the recognition of patterns of reasoning; challenging metaphysical assumption; and finally, theory building using construction cases.

Conclusion

In this paper we have argued that, once we distinguish the multifarious functions of cases, we can make best use of them through a specific sequence, which we call DRE. Though we think it has various advantages over traditional RE, we should emphasise that we do not see the model as rigid but fluid. For some problems or enquiries, some of these stages may be omitted (most obviously, construction cases may not be appropriate). Moreover, some movement back and forth between different stages in DRE is encouraged. For example, in the argumentative stage, movement back and forth between principles and argument cases (such as counterexamples or corner cases) may proceed in much the same manner as in RE.

[42] The term "intuition pump" was coined by Daniel Dennett in (1991), Allen Lane (ed.), Consciousness Explained, The Penguin Press.

Central to our development and defence of DRE has been a taxonomy of different cases and their functions. Oftentimes these different functions are already evident in the literature, even if they have not been identified explicitly. Some of these functions are also independent of the directional approach and can be employed in the course of RE or other case-based methodologies. Our aim has been to deepen, clarify and extend our understanding of these cases, rather then fully supplant previous methods. That said, we have also argued that, once the various functions of cases are properly itemised, they fit neatly into and complement our directional approach.