14 "Imagine If They Did That to You!"

The Complexity of Empathy

Luke Roelofs

We often remonstrate with each other by appealing to imagined role reversals. If I've done something that seems to you obviously wrong but to me seems fine or trivial, you might try and make me see the problem by having me imagine the same action done to me by someone else. Indeed, the idea of switching places with those our actions affect is central to the so-called "golden rule", to treat others as one would wish them to treat oneself, which is sometimes held up as a near-universally endorsed moral principle. But what is going on, cognitively, when we imagine such a role reversal? Why does imagining something that did not happen change my view of what did happen? What is the relationship between imagining this counterfactual scenario, and imagining the actual scenario from the other person's point of view? And how does either relate to "imagining being" the other person?

This question connects to a much bigger question about the place of empathy in a world increasingly threatened by climate change, a pandemic, and a rising global tide of fascism. It's common and popular to argue that we need more empathy. Shortly before Barack Obama's election, he wrote that the USA suffers from an "an empathy deficit", and that "a stronger sense of empathy would tilt the balance of our politics in favour of those who are struggling" (Obama 2006, pp. 67–8); halfway through his term, Ta-Nehisi Coates suggested that difficulty with the "basic extension of empathy is one of the great barriers in understanding race in this country" (Coates 2011). A recent book about why white Evangelicals support Trump is titled *The End of Empathy* (Compton 2020); a call to arms after Trump's inauguration laid out the hope for an "Empathy Politics" (Serano 2017), and it's always tempting, when trying to get middle-class White people to care about racism, to show them a video of police brutalising a Black person and ask "how would you feel?"

But this "empathy economy" (Stinson 2020) invites pushback; critics claim that "Antiracism is far less about empathy, about appealing to similarities, than it is about love, about honoring and protecting differences" (Salesses 2020). Even those who call for more empathy often feel a need to add caveats: Coates (2011) says "I do not mean a soft, flattering, hand-holding empathy. I mean a muscular empathy rooted in

curiosity", and Serano (2017) says "I am not talking about the shallow empathy of those who only seem capable of identifying with people who they perceive to be their own kind". I think this sort of qualified position is right: it's not simply that empathy is good, or that empathy is bad, but that empathy is *difficult* (cf. Epley and Caruso 2008; Langkau this volume; Kind this volume; and Jamison 2014, who writes "empathy is an edifice we build like a home or office – with architecture and design, scaffolding and electricity", p. 15). A key part of this difficulty is balancing the recognition of otherness with the search for commonality, and my aim in this chapter is to contribute to untangling some of the conceptual issues involved in finding this balance.

I will keep my focus on "empathic role reversal" without trying to define or delimit "empathy" as a general category. I argue that we need to separate two distinct questions about any given act of empathetic imagining, namely, "what mental states does it create?" and "what thing or possibility does it target?". When we do so, we can see how the various linguistic forms available ("imagine if", "imagine being", "imagine from their point of view", etc.) do not name distinct cognitive acts so much as they emphasise different steps of a complex cognitive process, which is primarily directed at the actual other, but often instrumentally directed at anything counterfactual. In doing so I hope to show why certain lines of criticism of empathy fail, and why empathic role reversal has value despite the complex difficulties of doing it well.

14.1 Empathic Role Reversal

To focus the discussion, consider case 1:

> **Case 1:** A has wronged B, but doesn't really care. C remonstrates with A, saying "how could you do that to B? Imagine if B did that to you! How would you feel?" A imagines that: they imagine B wronging A in the same way. They find themselves feeling outraged and hurt about B's imagined act. As a result, they revise their attitude to their own actions, and seek out B to make amends.

In this case, imagining a counterfactual scenario with roles reversed changes A's attitude and feelings about the actual world. Of course there is a large question here about how the final step works, where A changes their moral judgements and is moved to new actions. But my interest in the previous step, the imagining, and particularly in the apparent contrast with case 2:

> **Case 2:** A has wronged B, but doesn't really care. C remonstrates with A, saying "How could you do that to B? Look at this situation from B's perspective! Imagine how they must feel!" A imagines that:

> they imagine the situation from B's perspective, and find themselves feeling outraged and hurt about the actual action which they themselves have performed. As a result, they revise their attitude to their actions, and seek out B to make amends.

In this case, it might seem, no counterfactual scenario has been imagined, only the actual situation itself. Yet in some sense, cases 1 and 2 seem very similar; they both result in the same change of heart, and pre-theoretically it might seem strange to insist on a profound cognitive difference in what A does in the two cases.

But now finally consider a third case:

> **Case 3:** A has wronged B, but doesn't really care. C remonstrates with A, saying "How could you do that to B? Imagine if you were B in this situation." They follow up with one or the other of the following instructions: (a) "Imagine how they would feel! (b) Imagine how you would feel! (c) Imagine how that would feel!" A complies with the instructions given: they imagine being B, suffering the wrong they have themselves just inflicted. Finding themselves feeling outraged and hurt, they revise their attitude to their actions, and seek out B to make amends.

What is going on here? It seems very similar to cases 1 and 2: A is imaginatively "putting themselves into the shoes of" their victim, and finds that this exercise changes their feelings about their actions. But are they imagining the actual situation (where B is the victim, and A the agent), or a counterfactual situation (where A is the victim, and someone else the agent), or some metaphysically impossible situation (where A is B, and A is the victim, and B is the victim)?

It would be very natural to say that in case 3, A is "imaginatively identifying themselves with B". But then who is A identifying themselves with in case 2, if not B? Are they identifying themselves with *themselves* in case 1? Is case 3 actually just a more perspicuously described case 2? And in case 2, where A "imagines how B feels", are they necessarily thereby "imagining being B"?

I will argue that these three cases involve substantively the same mental process (what I have been calling "empathic role reversal"). They may differ in certain subsidiary respects, but these differences are only loosely tied to the different forms of language employed. In each case, C's aim is to for A to understand and be moved by B's experience, and towards this end they may direct A's attention either to actual or counterfactual scenarios, depending on their priorities. In other words, there is here no very sharp or deep distinction between "self-imagining" and "other-imagining". Many writers have claimed that some version of this distinction is of major importance: Gordon, for example, claims that

Adam Smith's *Theory of Moral Sentiments* is flawed because it runs the two together (1995a, cf. 1995b, p. 55, Nanay 2010); Batson claims to find empirical differences between the two, which he fears too much study has neglected (Batson et al. 1997; Batson 2009; cf. Stotland 1969); competing accounts of engagement with fictional settings and characters turn on the distinction (e.g. Currie 1995; Smith 1995, 1997; Gaut 1998; Giovannelli 2008); and many philosophers have wondered if there is something inherently impossible or misguided about one or the other (Williams 1973; cf. Wollheim 1984; Walton 1990; Velleman 1996; Wirling 2014; Langton 2019). In particular, Peter Goldie has argued that while self-imagining is a valid and useful cognitive tool, other-imagining is both conceptually impossible and ethically suspect (2011).

I do not want to claim that there is no distinction at all between "imagining being myself (in some situation)" and "imagining being the other". But I do deny that there is a single, clear, stable distinction to draw, that is significant across contexts. On the contrary, I argue, there are multiple distinctions here, which do not always line up, and which often show intermediate, indeterminate, or mixed forms. Moreover, I claim that in empathic role reversal, imaginers often employ multiple forms of imagination, at once or sequentially, criss-crossing these distinctions, in service of a single goal. As a result, theories like Smith's account of sympathy, which do not carefully distinguish between self-imagining and other-imagining, may be right not to do so, since for certain purposes that distinction obscures as much as it clarifies.

To be precise, I think there are at least three orthogonal distinctions in this vicinity. A full treatment of them will require the distinction between model and target that I will draw in Section 14.3, but they may be sketched briefly here.

First, there is what we may call the *identity distinction*, between imaginings that represent the imaginer, in some actual or possible circumstances, and those that represent another person (this is the standard way to distinguish "self-imagining" from "other-imagining").

Second, there is what we may call the *adjustment distinction*, between imaginings whose content matches the imaginer's own personality, character, and background, and those that make adjustments to simulate the personality, character, and background of another person.

Third, there is what we may call the *internal-external distinction*, between imaginings that take an "external" view on something and imaginings that are "from the inside", recreating the experience of a participant. It is the distinction between, for instance, simply imagining that some events occur, or imagining seeing them from some impersonal viewpoint, and imagining performing or suffering some active role in those events. This distinction is itself highly disputed, drawn in multiple ways and given different labels (such as "central" and "acentral",

"subjective" and "objective", "experiential" and "non-experiential"), so I will try not to lean too heavily on any particular account of it; for discussion see Wollheim (1984), Vendler (1984), Recanati (2007), Giovannelli (2008), and Dokic and Arcangeli (2015).

I will suggest, in Section 14.4, that keeping all three of these rough distinctions in mind will give us a better sense of how empathic role reversal functions. But first I want to say a bit more about Goldie's critique of empathy, because he articulates most sharply the idea that there is a large difference between self-imagining and other-imagining, and that it matters greatly which one we are involved in on a particular occasion.

14.2 Goldie's Critique of Empathy

Here is how Goldie puts his position:

> …what I am against is what I will call *empathetic perspective-shifting*: consciously and intentionally shifting your perspective in order to imagine *being* the other person, and thereby sharing in *his or her* thoughts, feelings, decisions, and other aspects of their psychology. I am not against what I will call *in-his-shoes perspective-shifting*: consciously and intentionally shifting your perspective in order to imagine what thoughts, feelings, decisions, and so on *you* would arrive at if you were in the other's circumstances. These two […] are often either confused or not sufficiently distinguished. What I want to do here is to show just how different they are, and how deep the problems are for empathetic perspective-shifting.
>
> (Goldie 2011, p. 302)

I think the distinction Goldie draws here (empathetic vs. in-their-shoes) is confused, running together two distinct questions. He draws his distinction both in terms of what sort of mental states you imagine having (the sort you would have vs. the sort the other actually has), and in terms of the identity of the subject imagined (whether it is you or them). These two dimensions, I will argue, come apart (they are what I above called the *identity distinction* and the *adjustment distinction*). Later on he re-asserts the latter, identity-based, version of the distinction:

> […] the difference between in-his-shoes perspective-shifting and empathetic perspective-shifting lies in the content of the imaginative project: who, *in the imaginative project*, is doing the thinking […]. [in] in-his-shoes perspective-taking […] A imagines *himself* in that situation […] [while in] empathetic perspective-shifting […] A imagines *being B* in that situation […]
>
> (Goldie 2011, p. 305)

Despite committing to this identity-based way of drawing the distinction, the objections Goldie raises revolve around the adjusted or unadjusted character of the mental states one imagines having. Indeed, he admits that in cases where A and B are psychologically similar enough, in relevant respects, and not relevantly irrational, the two forms of perspective-shifting "will each produce the same result, the same 'output'" (Goldie 2011, p. 307). Consequently, he thinks, there will be no detectable problem with empathetic perspective-shifting in such cases. The inherent problems with it become manifest only in more "ambitious" cases, in particular in cases where A and B have importantly different characters, so that what A would think and feel in a given situation diverges from what B would think and feel. (This also applies in cases where irrational or unconscious factors strongly influence B's thought process, so that what "any rational agent" would think and feel in a given situation diverges from what B would think and feel.) In both cases, he insists, A cannot successfully and accurately imagine being B, and even if A recognises these differences, and tries to adjust their imagining to incorporate them, the very fact that they must consciously and explicitly make those adjustments will "falsify" the character of B's perspective, for whom they are not explicit.

Of course, A might try to imagine what A themselves would think and feel if they were (i) in B's situation, and (ii) had B's psychological character – this would seemingly be in-their-shoes perspective-shifting, but with all the same problems as empathetic perspective-shifting. Goldie admits this, but says that:

> […]at this point, this more ambitious variety of in-his-shoes perspective-shifting begins to suffer from exactly the same problems as empathetic perspective-shifting[…] the lesson from this for in-his-shoes perspective-shifting, I think, is that it should not be too ambitious.
>
> (Goldie 2011, p. 312)

I think some good direct answers to Goldie's critique are available; Langkau (this volume) presents a number (see also Coeckelbergh 2007). My aim here is not to directly defend empathetic perspective-taking, but to challenge the framework within which Goldie makes the critique: that there are two distinct cognitive acts here, that we have to choose which to engage in, and that it makes sense to be "against" one but not the other. On the contrary, the identity of the person imagined is a separate question from the adjusted or unadjusted character of the mental states imagined – and Goldie's own discussions show the cracks on this score, with the need to restrict his objections to "ambitious cases" rather than base cases, and to admit that even his favoured form of perspective-shifting, the "in-their-shoes" form, is problematic in

ambitious cases. The force of his critique concerns not the identity of the person imagined but the amount of psychological adjustment attempted, and so whatever force there is to the critique (which, as noted, I think Langkau addresses quite effectively) it affects "in-their-shoes" and "empathetic" perspective-shifting equally. Moreover, I hope to show, neither the identity distinction nor the character distinction is sharp; many cases fall somewhere in between. And this intermediate character is often not just a result of sloppiness or inattention, but a positive part of the skill of empathising well.

14.3 Two Readings of "What Do I Imagine?"

I claim there is a systematic ambiguity in the way we describe imagination's *objects*: there are two very different ways of asking and answering the question "what is imagined?" Roughly, this question can either ask what states (images, feelings, propositional thoughts, etc.) are created in the imaginer's mind or ask what external thing the imaginer thereby represents. I will distinguish these by talking about what an imaginer mentally *creates*, and what their imagining *targets*. Note that this is not a matter of there being two senses of imagination, two different mental processes: it is two aspects of a single process, which a single linguistic form (using "imagine" as a verb with a direct object) ambiguously expresses.

To see what this distinction comes to, consider some simple examples. I might form in my mind a visual image as of a soft-bodied creature with two eyes, an elongated conical body, and ten limbs. This might be as part of a few different imaginative projects: I might be trying (badly) to imagine what octopuses look like, or to imagine what the octopus I will meet tomorrow (when I visit the zoo) will look like, or to imagine what I would look like if I were turned into an octopus. These are three very different targets: one is (roughly) a general category, one is a specific individual, and one is a hypothetical version of myself. In that sense, "what I imagine" is very different. Yet in another sense, "what I imagine" is in each case the same: I imagine this image. If someone asked me "what are you imagining?", they might be asking about either. I would say: the image is what I create, while the general category, specific individual, or hypothetical version of myself is what I target.

One notable thing is that the image may be evaluated for accuracy, but the standard for accuracy is set by the target: if the octopus I will meet tomorrow looks very unusual, then an image that would be acceptably accurate as "what octopuses look like" might be importantly inaccurate as "what the octopus I will meet tomorrow will look like". And "what I would look like if I were turned into an octopus" is arguably not definite enough to give much in the way of accuracy conditions. As a matter of fact, of course, this image would be seriously inaccurate in any of these three cases: its shape is distinctly un-octopuslike, because

the description I gave above is a description of a squid, not an octopus. Nevertheless, if my intention is to imagine an octopus, the image I form is an inaccurate image of an octopus – not an accurate image of a squid. But if I had instead been aiming to imagine what squid look like, I might have formed the same image, and then it would have been an accurate image of a squid.

It is, I think, easier to present this distinction with regard to visual images, since it is fairly clear that the image by itself under-determines the target. But I think the same point extends to other forms of imagining: including, in particular, the imagining of thoughts, feelings, decisions, and so on that we are interested in here. When A imagines being wronged by B, they will recreate in their mind a certain complex of connected thoughts, images, emotions, wishes, assumptions, dispositions, etc. It would be very odd to call this complex (the thing that they mentally create) an "image", but I am unsure what exactly to call it. The best term I have found is "model": on my account, imagining involves creating in my mind a model, and using it to represent (to target) something outside my mind (the target).

This account is meant to be neutral on how to analyse the states involved in imagining. I am most attracted to what is sometimes called a "recreativist" or "simulationist" account of the imagination (for discussion see Currie and Ravenscroft 2002; Goldman 2006; Tagliafico 2011; Kind 2013; Balcerak Jackson 2018), on which it is essentially the capacity to create and manipulate "offline versions" of other mental states. A mental image is an offline copy of perception, and thus just one species of the general class of imaginings. Then we would say that whenever we *create* an imaginative model, it is formed of offline states which are *recreations* of other, online, mental state types. Recreativism is congenial to my analysis because it offers a unified account of the sort of mental states that form the model (they are "offline copies"), but other accounts, including those with a more heterogeneous account of the states involved (e.g. Dorsch 2012; Kind 2013; Langland-Hassan 2020), are also compatible with the distinction drawn here.

What we mentally create is, by definition, some set of mental states. By contrast, what we target is, I think, more or less unlimited (though there may be targets which we cannot imagine accurately or usefully). We can target definite scenarios like "what will happen", or "what it would have looked like", or "how she must feel now", or "what it would sound like if…", but we can also target generic possibilities ("what would a human-sized frog look like?"), and our targets can be governed by the rules of a game ("pretend that I'm a human-sized frog") or a public fiction ("once upon a time there was a human-sized frog").

My distinction closely resembles the distinction common in discussions of pictorial representation (e.g. Abell 2005, 2007; Greenberg 2018), between a picture's content (what it says) and its object or target

(what it is a picture of). The former is what can be judged accurate or inaccurate, but the latter is what provides the standard of accuracy. My aim is to extend this distinction to apply not just to physical pictures and mental images, but to the full range of imaginings (an extension which Greenberg suggests but does not develop, 2018, p. 895). Models express contents, relative to the context (especially the intentions) surrounding their creation, and those contexts also specify a target relative to which the model is accurate if its content holds.

Following Greenberg, I deny that targets can be identified with any element of content (such as "singular" content as opposed to "attributive content"). Likewise, my model-target distinction does not directly map onto Currie and Ravenscroft's (2002, p. 12 ff) distinction between an imagining's "content" and its "character". They draw this distinction primarily to insist that while visual imagining, for instance, always has a visual "character", it does not therefore have "seeing" as part of its content. I very much agree with this point, and in particular with the thought that the phenomenal character of an imagining does not dictate its target, but the content-character distinction they draw does not directly match my target-model distinction. More broadly, I do not think the model-target distinction can be straightforwardly identified with familiar distinctions like attitude-content, sense-reference, or predicate-subject. This is for the simple reason that imaginative models can contain recreations of the states within which we draw all of these distinctions. For example, hoping that Hesperus is Phosphorus and imagining hoping that involves the same attitude (hoping), content (that Hesperus is Phosphorus), subject (Hesperus), predicate (being Phosphorus), senses ("Hesperus", "Phosphorus"), and referent (the planet Venus). There may be many interesting connections between model-target and these other distinctions, but they are not simply the same.

Although recreation and targeting work together, their connection is quite loose: as my above examples illustrated, the same model can stand in for many targets, and the same object can be targeted using many different models. Moreover, not all aspects of what we recreate need to be representing anything about the target. For example, I might tell you about the shape of a country by asking you to "imagine a hexagon", and telling you that the country is roughly that shape. What you create is an image of a straight-edged regular hexagon (maybe a blue one!), but the straightness of the edges (and the blueness) is irrelevant to what you are targeting (the approximate shape of the country), and is not to be evaluated for accuracy.

We might wonder at this point about "idle" conscious intentions. Since there is no obvious way for such imaginings to be inaccurate, we might think they have no target: this would be perfectly compatible with the account I am offering. Not all imaginings need to have targets. Alternatively, we might think of them as targeting something like "objects

that look like this", so that they are always trivially accurate. This would have the advantage of allowing idle imaginings to constrain what counts as accuracy for any further thoughts that aim to "build on" them.

I think that recreating and targeting can be varied independently of each other more or less however we want, at least when we form explicit intentions. However, in practice we often do not, and then the two intertwine because our minds (fortunately) tend to work in ways that tend towards accuracy. If I set out to imagine an octopus, I will usually form an image that would be fairly accurate, and if I form an image that is inaccurate (like the squid-image above), it will usually not be wildly inaccurate (not an image of the empire state building, for instance). Conversely, if an octopus-like image pops into my head unbidden, I am likely to take it as standing in for octopuses, and not for American architecture. By "I am likely to take it as" standing in for (i.e. targetting) octopuses, I mean two things: first, that if I do have any explicit thoughts about "what it is" (i.e. about what I represent by forming this image), those explicit thoughts are likely to say "it's an octopus", and second, that even if I never think explicitly about "what it is", I am likely to "proceed from it" (i.e. draw inferences, form further images, have thoughts, dredge up memories) in ways that pertain to octopuses, not to American architecture. Plausibly, in the second sort of case (where there is no explicit thought about what I am targetting), these functional facts about how I proceed (what role the image plays in my ongoing mental life) are what constitute it being about, i.e. targetting, one thing rather than another.

Because of the psychological and constitutive facts just canvassed, it will often be sensible to assume that an image (or other sort of imaginative model) targets *whatever it would be most accurate* in targeting, and thus that what I imaginatively target is fixed by what I imaginatively recreate. But, I claim, this is just a rough-and-ready heuristic, and the two imaginative relations have no fixed correspondence.

14.4 The Complexity of Empathy

Using this distinction between model and target, let us return to the three cases in Section 14.1. I said there that I think all three involve substantively the same, complex, cognitive process: the differences in language used reflect differences in the presence or prominence of different parts of this process, but not differences in what is, in the first instance, going on. Now I can explain this claim of substantive sameness more fully.

14.4.1 Building Models and Shifting Targets

My basic claim is that cases 1 and 2 can involve different targets (a counterfactual situation, and the other's perspective on the actual situation) but the very same model. They involve the same model insofar as they

all recreate the same suite of connected mental states, namely those that are typically involved in suffering the kind of wrong done by A. B actually underwent that experience, in the counterfactual scenario where A is the victim, A has that experience. The model matches both, but just like the same image can represent different sights, this model could be used to represent B's actual experience or A's possible experience with equal accuracy. Of course in practice there may be lots of variation in the model created: A might include or emphasise different details in case 1 that they do not in case 2 (see below for fuller discussion), but there is no *necessary* difference in model between the two cases. In that sense, A "imagines the same thing" in both cases.

I said that this shared model was being used to represent two different targets, but that is not the whole story. It is true that in case 1, it seems that A sets out to comply with C's instruction to imagine a counterfactual ("if they did that to you"), and so we should take their intention as fixing their target: the counterfactual scenario. No such counterfactual is mentioned in case 2, where A seems to intentionally target the actual situation. But despite this difference in *initial* targets, I think that in both cases A *ends up* targeting B's perspective on the actual situation. In case 1 they just come at that target more indirectly, first recreating the suite of mental states typical of being a victim of that kind of wrong (organised by the intention to imagine the counterfactual scenario), but then keeping that model while shifting its target to be the actual situation. This shift, I think, is easy precisely because there is no change in the model, but only in what the subject takes it to stand in for. To motivate the thought that there is this final target-shift, contrast case 1 with case 4:

Case 4: A has wronged B, but doesn't really care. C remonstrates with A, saying "how could you do that to B? Imagine if B did that to you! How would you feel?" A imagines that: they imagine B wronging A in the same way. They find themselves feeling outraged and hurt about "what B has done". They turn to C and say "you're right! Thank goodness B didn't do that to me, that would be awful. I'm so relieved that this is just a counterfactual, it really helps me appreciate my actual situation more by contrast".

I think A in this case has complied with the *letter* of C's instruction perfectly well: they have imagined if B did that to them, just as accurately and insightfully as in case 1. But they have clearly not done everything that C hoped for, because they have not changed their attitude towards B's actual situation. C's aim (and A's aim insofar as A is trying to go along with C) is to use the counterfactual to sharpen A's grasp of the actual situation, and when that does not happen, C's remonstrance has failed, even if A successfully imagines the counterfactual.

All I will insist on is that the overall aim and target of the imagining is the same in cases 1 and 2, with the difference in immediate targets being subsidiary to this. But a stronger claim could also be made: that at least sometimes, the imagining process involved cannot be cleanly divided between self-imagining and other-imagining. I suggested above that what a given imagining targets might be fixed partly by functional facts about what further role it plays in the subject's mind. If A, as soon as they form this model (i.e. connected offline versions of the states typical of being wronged), is disposed both to infer things from it about a fanciful counterfactual (e.g. to worry they would not respond gracefully, or to resolve to take precautions against it) and to change their attitude to the actual B (e.g. to feel regret and shame, and resolve to apologise), then these dispositions might support thinking that they target two different targets at once with the same model, or that what they target with that model is indeterminate, or (if the dispositions develop at different rates) that they targeted first one thing and then another, but without any determinate moment when the shift happened. And arguably this is implicit in C's instruction in case 1: case 4 illustrates that *simply* imagining the counterfactual scenario is not really complying with what C clearly meant. So if A's target is fixed by their intention, and their intention is to follow C's instruction, then their target again might be *both* self and other.

So the first of the three distinctions I noted in Section 14.1, regarding the identity of the person imagined, is not always in practice sharp. It is, I think, easy to show that the other two are likewise prone to intermediate forms. With the issue of psychological adjustment, it is clear that more or less adjustment may be made: there are a hundred ways in which any two people differ, and even if I make some adjustments to my imagining to reflect some of these differences, there are more that I could make. And although the distinction between imagining "from the outside" and "from the inside" is formally binary, the same imagining may be from inside some objects but from outside others. In Vendler's famous example, when I imagine swimming in the sea "from the outside", i.e. imagine seeing myself swimming, I may still be imagining "looking down at the sea from a cliff" from the inside. So without fixing which particular feature we are talking about, we will not be able to exhaustively distinguish imaginings "from the inside" and "from the outside".

So the very same, dual-targetted, imagining could be appropriate for both case 1 and case 2. This point is reinforced when we turn to consider case 3. I don't think this case really involves a different mental process than cases 1 and 2; rather, I think the linguistic form used here ("imagine if you were B") is just a tool for efficiently gesturing towards the same overall mental process as in cases 1 and 2: recreating the experience of victimhood and targeting it at B's actual experience, perhaps by first setting out to target a self-involving counterfactual. The seemingly impossible counterfactual of "if A were B" is a red herring: the point is to

convey concisely that it is *B* whose perspective one is concerned with, but that what one should imagine is their first-person perspective.

14.4.2 Imagining From the Inside and From the Outside

The variety of idioms here, and the complex connections of ultimate and immediate targets, might seem messy and confused. But I think they are actually functional: different forms of language can make salient different challenges and risks involved in empathic role reversal. Consider first the question: why bother with the self-involving counterfactual? Why not simply tell A to imagine how B feels? I think we can see the value of the detour through self-imagining by considering another case where the remonstrance fails:

Case 5: A has wronged B, but doesn't really care. C remonstrates with A, saying "How could you do that to B? Imagine doing that to B, how do you think they feel?" A, a little confused, dutifully imagines doing what they did to B, and imagines that B feels hurt and outraged. But they are not moved; instead they reply: "I don't have to imagine it, I know it already: I did that, and yeah B is mad about it, but whatever. Imagining the actual situation isn't going to change how I feel about it".

Like in case 2, here the only situation targeted is the actual situation; yet there is seemingly a big difference in *how* it is imagined. Case 2 is more like case 1 than like case 5, because in cases 1 and 2, the model constructed is "from the victim's perspective", i.e. it recreates the sorts of mental states a victim would typically have (beliefs, perceptions, memories, and then frustration, hurt, a sense of betrayal, etc.). But in case 5, A is clearly using a very different model – one that imagines B's experience "from the outside", to use the third distinction I noted in Section 14.1. They might be using a model that is "from the perpetrator's perspective", i.e. that recreates the same kinds of mental states that A already went through, when they committed the wrong in the first place. Alternatively, they might be using a model that is "from a third-person perspective", recreating little more than beliefs about what in fact happened. In either case, since they remember performing the action, and already know it happened, there is nothing new for them to learn from this imagining.

This sort of imaginative model is clearly ill-suited to accomplish what B is trying to accomplish. And the formation of models like this is arguably a major problem when trying to cultivate empathy. If A simply thinks about the actual case, about which they already have many beliefs, feelings, memories, etc., they might struggle to avoid forming a model that is contaminated by their own perspective – that privileges

how they see things and thereby obscures how B sees things. Their natural, self-involved, view on their own life forms a barrier to whatever in B's perspective might be uncomfortable or unfamiliar.

The value of the detour through self-imagining in case 1 is to help bypass this barrier. By turning aside from the actual world, A is better able to form a model that really matches the experience of suffering, as opposed to committing, the wrong in question. Their natural self-involvement is recruited to subserve empathy, rather than standing in its way. By starting with that self-referential intention, they cannot help but imagine "from the victim's point of view", i.e. what they recreate will be the vivid, outraged, experiences of a victim, and not the bored experiences of a perpetrator, or the bloodless belief that such and such has taken place. And when that model is in place, the remonstrator's hope might run, the shift in target is not so hard.

14.4.3 The Risk of Egocentric Bias

But there is an opposite risk, which a detour through self-imagining exacerbates rather than mitigates, namely the risk of "egocentric bias", of forming an imaginative model that reflects the peculiarities of one's own character, or fails to reflect the peculiarities of the other's character. This risk is at the centre of Goldie's critique of empathetic perspective-shifting: he ties the risk of imagining the other too much like ourselves, or at least not enough like the other themselves, specifically to other-imagining, which takes the other as its target. But the degree of egocentrism is a matter of the particular psychological states that form the imaginative model, which is dissociable from the imaginative target. To see this, consider final two cases:

Case 6: A has wronged B, but doesn't really care. C remonstrates with A, saying "how could you do that to B? Imagine if B did that to you! How would you feel?" At first, A imagines that: they imagine B wronging A in the same way. But they find themselves feeling no outrage or hurt, but merely a mild annoyance. When they report this, C objects: "that's because you and B are such different people: here are some ways that B's background and character make what you did especially bad…" After hearing these differences pointed out, A tries again, imagining themselves not only suffering the same actions as B did but also having the same relevant background and character as B. The hurt and outrage are now vivid, and A realises how badly they've acted.

Over the course of the conversation, A changes their model in such a way as to make it a more accurate model of B's actual perspective. Both before and after, the explicit target of their imagining is a self-involving

counterfactual scenario, just a more remote counterfactual in the latter case. But this fact is not the important one, as we see if we compare case 6 with the very similar case 7:

Case 7: A has wronged B, but doesn't really care. C remonstrates with A, saying "how could you do that to B? Imagine if you were B! How would that feel?" (a) A imagines that: they imagine being B, suffering this action performed by A. But they find themselves feeling no outrage or hurt, but merely a mild annoyance. (b) When they report this, C objects: "that's because you and B are such different people: here are some ways that B's background and character make what you did especially bad…" After hearing these differences pointed out, A tries again, imagining being B, but now with a more accurate model that incorporates relevant background and character. The hurt and outrage are now vivid, and A realises how badly they've acted.

Here the change in model, to reduce its egocentrism, has nothing to do with a change in target; it is just about getting a more accurate model of the same target. But it seems to me that what A is doing here is still the same complex activity as in case 6 (just like cases 1, 2, and 3 all involve the same complex activity), and that the ultimate point in both cases is to end with the right kind of model for understanding, and being moved by, B's actual experience. The diverse explicit targets, like the counterfactual where A has different character traits, are only instrumental tools for helping A to construct that model.

Telling A specifically to imagine *B*'s perspective may be a way to make salient the risk of egocentric bias. Conversely, telling A to imagine *themselves* in B's situation may be a way to make salient the opposite risk, of imagining what happened only "from the outside", and fail to escape from one's existing view of the events. To pair with "egocentric bias", we might term this "allocentric indifference". The point is that empathic role reversal can go wrong in at least two opposite ways: the imaginer might fail to construct a model that captures the relevant type of experience "from the inside" (as in case 5), succumbing to allocentric indifference, or they might fail to construct an accurate model, by incorporating their own character and not the relevantly different character of the other (as in cases 6 and 7), succumbing to egocentric bias. Different linguistic forms serve to highlight and guard against different such failings: "imagine if it happened to *you*" emphasises that the model constructed should involve the kinds of states associated with suffering the wrong, not with committing it or simply knowing about it; "imagine if you were *them*" emphasises that the model constructed should reflect the specific characteristics of the other person; "imagine how they must (actually) feel" emphasises that the model is ultimately meant to target

something actual, not just a counterfactual. But it would be a mistake, I think, to fix on these different linguistic forms, and take them to be calling for distinct cognitive acts, which could then be compared and critiqued relative to one another. The act called for is grasping the other's situation from their perspective, which may require first targetting other scenarios in order to get the right model: this act is complex, and the same in each case.

But all of these challenges are what Langkau calls "psychological" and "epistemic" challenges (this volume, cf. Epley and Caruso 2009, compares the more general challenges of imagining skilfully discussed in Kind 2018, 2020): they involve things which are hard for us to do, or which we may lack the necessary knowledge to do. They are not, as Langkau puts it, "conceptual challenges", inherent problems with the very possibility of empathic role reversal, or with any particular form of it.

14.5 Conclusions

Of course my "build the model, then shift the target" account is not the only way to theorise empathic role reversal. In particular, one might naturally think of C's aim as being for A to learn a general truth from considering the counterfactual. For example, Nagel writes:

> If something I do will cause another creature to suffer, that counts against doing it. I can come to see that this is true by generalizing from the evident disvalue of my own suffering.
>
> (Nagel 2012, p. 77)

If A learns a general principle about the disvalue of some experience from self-imagining, they could then apply that general principle to their actual situation: the principle is what mediates between imagining the counterfactual and grasping the actual. I have not here refuted or objected to that account, and I think we often do draw general lessons from our imaginings, and apply them to new cases. But I think sometimes we also perform the more direct target-shifting process I have described, and having that account in view is useful for, I believe, a few reasons. It underlines how exceedingly close the processes in cases 1, 2, and 3 may be, and thereby explains why we often feel that it makes little difference which of the three linguistic forms we use. The target-shift account also, I think, better captures the concrete, first-personal character of the cognition in these cases – A does not need to try and formulate a general principle that explicitly captures all the relevant similarities between the counterfactual and actual scenarios: they can simply recognise the two scenarios as similar. And if we thought that imagining things from others' perspectives was of foundational importance in morality (as on the

views of Smith 1976; Nichols 2004; and others), then we might think that the general-principle account puts the cart before the horse: we arrive at general moral principles by abstracting from empathy with others, rather than relying on them to enable it.

I would add that the complexity of empathic role reversal, and the diversity of seemingly conflicting terminology for describing it, reflects the difficulty of empathising well. Empathy is a skill we have to keep learning, an ongoing balancing act between drawing the other too close, substituting an image of them modelled after ourselves, and letting them drift so far away that we lose sight of their humanity. Trying to understand others through imagination is, I believe, vital and noble, but it needs to be done in the keen awareness of how limited that understanding will remain, and with a resolve not to let our compassion be restricted by differences that challenge it.

References

Abell, Catharine. 2005. "Against Depictive Conventionalism." *American Philosophical Quarterly* 42(3), 185–97.

Abell, Catharine. 2007. "Pictorial Realism." *Australasian Journal of Philosophy* 85(1), 1–17.

Balcerak Jackson, Magdalena. 2018. "Justification by Imagination." In *Perceptual Memory and Perceptual Imagination*, edited by F. Dorsch and F. Macpherson, 209–26. Oxford: OUP.

Batson, C.D. 2009. "Two Forms of Perspective Taking: Imagining How Another Feels and Imagining How You Would Feel." In *Handbook of Imagination and Mental Simulation*, edited by K.D. Markman, W.M.P. Klein, and J.A. Suhr, 267–79. New York: Psychology Press.

Batson, C.D., Early, S., and Salvarani, G. 1997. "Perspective Taking: Imagining How Another Feels versus Imagining How You Would Feel." *Personality and Social Psychology Bulletin* 23, 751–8.

Coates, Ta-Nehisi. 2011. "A Muscular Empathy." *The Atlantic*. https://www.theatlantic.com/national/archive/2011/12/a-muscular-empathy/249984/

Coeckelbergh, Mark. 2007. "Who Needs Empathy? A Response to Goldie's Arguments against Empathy and Suggestions for an Account of Mutual Perspective-shifting in Contexts of Helping and Care." *Ethics and Education* 2(1), 61–72.

Compton, John. (2020). *The End of Empathy: Why White Protestants Stopped Loving Their Neighbors*. Oxford: Oxford University Press.

Currie, Gregory. 1995. *Image and Mind: Film, Philosophy and Cognitive Science*. Cambridge: Cambridge University Press.

Currie, Gregory, and Ravenscroft, Iain. 2002. *Recreative Minds: Imagination in Philosophy and Psychology*. Oxford: Clarendon Press.

Dokic, Jérôme, and Arcangeli, Margherita. 2015. "The Heterogeneity of Experiential Imagination." In *Open MIND. Philosophy and the Mind Sciences in the 21st Century*, edited by T.K. Metzinger and J.M. Windt, 431–50. Cambridge, MA: MIT Press.

Dorsch, Fabian. 2012. *The Unity of Imagining*. Germany: De Gruyter.
Epley, Nicholas and Caruso, Eugene M. 2009. "Perspective Taking: Misstepping into Others' Shoes." In *Handbook of Imagination and Mental Simulation*, edited by K.D. Markman, W.M.P. Klein, and J.A. Suhr, 295–309. New York: Psychology Press.
Giovannelli, Alessandro. 2008. "In and Out: The Dynamics of Imagination in the Engagement with Narratives." *Journal of Aesthetics and Art Criticism* 66(1), 11–24.
Goldie, Peter. 2011. "Anti-Empathy." In *Empathy: Philosophical and Psychological Perspectives*, edited by P. Goldie and A. Coplan, 302–18. Oxford: Oxford University Press.
Goldman, Alvin. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.
Gordon, R.M. 1995a. "Sympathy, Simulation, and the Impartial Spectator." *Ethics* 105, 727–42.
Gordon, Robert. 1995b. "Simulation without Introspection or Inference from Me to You." In *Mental Simulation: Evaluations and Applications - Reading in Mind and Language*, edited by Martin Davies and Tony Stone, 352–66. United Kingdom: Wiley.
Greenberg, Gabriel. 2018. "Content and Target in Pictorial Representation." *Ergo 5*, 865–98.
Gaut, Berys. 1998. "Imagination, Interpretation, and Film." *Philosophical Studies* 89, 331–41.
Jamison, Leslie. 2014. *The Empathy Exams: Essays*. United States: Graywolf Press.
Kind, Amy. 2013. "The Heterogeneity of the Imagination." *Erkenntnis* 78(1), 141–59.
Kind, Amy. 2018. "How Imagination Gives Rise to Knowledge." In *Perceptual Memory and Perceptual Imagination*, edited by Fabian Dorsch and Fiona Macpherson, 227–46. Oxford: Oxford University Press.
Kind, Amy. 2020. "What Imagination Teaches." In *Becoming Someone New: Essays on Transformative Experience, Choice, and Change*, edited by J. Schwenkler and E. Lambert, 133–46. Oxford: Oxford University Press.
Langland-Hassan, Peter. 2020. *Explaining Imagination*. Oxford: Oxford University Press.
Langton, Rae. 2019. "Empathy and First-Personal Imagining." *Proceedings of the Aristotelian Society* 69(1), 77–104.
Nagel, Thomas. 2012. *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False*. Oxford: Oxford University Press.
Nanay, Bence. 2010. "Adam Smith's Concept of Sympathy and its Contemporary Interpretations." *Adam Smith Review 5*, 85–105.
Nichols, Shaun. 2004. "Review of Currie and Ravenscroft, Recreative Minds." *Mind* 113, 329–34.
Obama, Barack. 2006. *The Audacity of Hope: Thoughts on Reclaiming the American Dream*. New York: Crown Press.
Recanati, François. 2007. "Imagining de se." https://jeannicod.ccsd.cnrs.fr/ijn_00160757

Salesses, Matthew. 2020. "The Empathy Economy Is a Sham. The Protest Movement Is Real." *The Daily Beast*. https://www.thedailybeast.com/the-empathy-economy-is-a-sham-the-protest-movement-is-real

Serano, Julia. 2017. Empathy Politics. Medium.com. https://medium.com/@juliaserano/empathy-politics-d7f62aa90e75

Smith, A. 1976. *The Theory of Moral Sentiments*, D.D. Raphael and A.L Macfie (eds), Oxford: Clarendon Press.

Smith, M. 1995. *Engaging Characters*. Oxford: Oxford University Press.

Smith, M. 1997. "Imagining from the Inside." In *Film Theory and Philosophy*, edited by R. Allen and M. Smith, 412–30. Oxford: Oxford University Press.

Stinson, Liz. 2020. "The Empathy Economy Is Booming, but What Happens When Our Emotional Connections to Others Are Designed, Packaged, and Sold?" *AIGA Eye of Design*. https://eyeondesign.aiga.org/the-empathy-economy-is-booming-but-what-happens-when-our-emotional-connections-to-others-are-designed-packaged-and-sold/

Stotland, E. 1969. "Exploratory Investigations of Empathy." In *Advances in Experimental Social Psychology*, edited by L. Berkowitz, vol. 4, 271–313. New York: Academic Press.

Tagliafico, Daniela. 2011. "Can we really speak of 'pretend desire'?" *Logic and Philosophy of Science* 9(1), 461–67.

Vendler, Zeno. 1984. *The Matter of Minds*. Oxford: Clarendon Press.

Velleman, David. 1996. "Self to Self." *The Philosophical Review* 105(1), 39–76.

Walton, Kendall. 1990. *Mimesis and Make-Believe: On the Foundations of the Representational Arts*. Cambridge, MA: Harvard University Press.

Williams, Bernard. 1973. "Imagination and the Self." In *Problems of the Self*, edited by B. Williams, 26–45. Cambridge: Cambridge University Press.

Wirling, Ylwa. 2014. "Imagining Oneself Being Someone Else – The Role of the Self in the Shoes of Another." *Journal of Consciousness Studies* 21(9–10), 205–25.

Wollheim, Richard. 1984. *The Thread of Life*. Cambridge, MA: Harvard University Press.