Ethical Theory and Moral Practice (2024) 27:337–352
https://doi.org/10.1007/s10677-023-10421-0

The Moral Permissibility of Perspective-Taking Interventions

Hannah Read[1] · **Thomas Douglas**[2]

Accepted: 4 October 2023 / Published online: 16 November 2023

Abstract
Interventions designed to promote perspective taking are increasingly prevalent in educational settings, and are also being considered for applications in other domains. Thus far, these perspective-taking interventions (PTIs) have largely escaped philosophical attention, however they are sometimes *prima facie* morally problematic in at least two respects: they are neither transparent nor easy to resist. Nontransparent or hard-to-resist PTIs call for a moral defense and our primary aim in this paper is to provide such a defense. We offer two arguments for the view that an exemplar PTI is morally permissible even though it is plausibly neither transparent nor easy to resist. The first argument appeals to an analogy between PTIs and permissible deceptive research practices. The second appeals to the way in which PTIs draw participants' attention to their reasons for action. We also respond to the objection that, by imposing a particular conception of the good, PTIs violate liberal neutrality.

Keywords Perspective taking · Moral permissibility · Nudging · Politics

1 Introduction

Perspective-taking interventions (PTIs) are socio-behavioral interventions that aim to improve individuals' abilities to take others' perspectives. These interventions typically involve altering the choice environment in ways that encourage individuals to imagine what it must be like for others in their situation, or what it would be like to be in another person's situation (Batson 2012). Perspective taking involves imaginatively projecting oneself into another person's situation (Batson et al. 1997; Batson 2012; Coplan 2011) and has been shown to promote a range of positive outcomes. These include improving interpersonal and intergroup relations, reducing bias and prejudice, and blurring harmful perceptions of group boundaries (Todd and Galinsky 2014; Gutsell et al. 2020; Simonovits et al. 2018; Gehlbach et al. 2015; Galinsky and Moskowitz 2000; Myers et al. 2014). Individuals are

✉ Thomas Douglas
thomas.douglas@philosophy.ox.ac.uk

Hannah Read
hannahread01@gmail.com

[1] Duke University, Durham, NC 27708, USA

[2] Oxford University, LittleGate House, 16-17 Saint Ebbe's St, Oxford OX1 1PT, UK

often unmotivated to practice and develop perspective-taking abilities when it comes to the perspectives of those outside their group, which limits the social benefits that perspective taking can yield (Davis and Maitner 2009).

PTIs are important because they can promote perspective taking in the face of these limitations. To date, PTIs have chiefly been employed in educational contexts—for instance, as part of cooperative learning and teaching models (Aronson 2002; Bridgeman 1981; Walker and Crogan 1998). These PTIs involve group activities where perspective taking is needed in order to succeed at the task at hand—for example, when students must take others perspectives in order to effectively convey information to other group members and thus enable the group to succeed on a task. We return to discuss this sort of PTI in greater detail below. However, PTIs have also been used in efforts to reduce prejudice outside the classroom (Todd and Galinsky 2014)—as with Broockman and Kalla's (2016) innovative study of the effects of encouraging perspective taking on antitransgender prejudice in political contexts—and may be more widely implemented in the future, for example as part of deliberative democratic processes (Muradova 2021; Read 2023) or to aid restorative justice efforts (Bloch 2021).

Yet, despite being widely and increasingly employed (Greenberg et al. 2017; Weissberg 2019), and having attracted little critical scrutiny, PTIs turn out to be more complicated to justify than initially appears. This is because, as we show below, PTIs are sometimes non-transparent and hard to resist. The moral permissibility of such PTIs calls for a moral defense and our aim in this paper is to provide such a defense.

We maintain that PTIs can be permissible even when they are neither transparent nor easy to resist. We offer an example of one such PTI that is neither transparent nor easy to resist and develop two arguments in support of the claim that this PTI is permissible. The first appeals to an analogy between the PTI and morally permissible forms of deception used for research purposes, and the second to an analogy between the PTI and certain nudges which operate by drawing the nudgee's attention to their reasons for action. These analogies are especially helpful given the significant amount of attention that has been paid to deceptive research and the ethics of nudging and the fact that the moral permissibility of PTIs has gone mostly unassessed.

The paper proceeds as follows. We begin, in Section 2, by introducing a paradigmatic but hypothetical PTI which will serve as our exemplar throughout. Next, in Section 3, we introduce two grounds for doubting the moral permissibility of some PTIs, including our exemplar. In Sections 4 and 5 we offer our two defenses of our exemplar PTI, before, in Section 6, responding to an important objection to these.

Before proceeding to our argument, a clarification is in order. At several points in our discussion, we draw on the literature on the moral permissibility of nudges—which we take to be interventions that (a) alter an individual's environment, (b) in order to influence her decisions, (c) by harnessing rapid, low-effort decision-making heuristics, such as 'choose what is most salient', 'stick with the default' or 'listen to people you recognise', and (d) not by employing incentives. Familiar examples of nudges include placing healthy foods in prominent locations in a cafeteria to encourage healthy food choices (Arno and Thomas 2016) and introducing an opt-out system for organ donation to increase donation rates (Rithalia et al. 2009). In drawing on this literature, we do not mean to imply that PTIs *are* nudges. Indeed, there are important differences between nudges and PTIs. Most notably, nudges are often defined such that they are easy to resist (Thaler and Sunstein 2009; Saghai 2013). By contrast, some PTIs, including the exemplar that will be our focus in this paper, fail to meet this criterion. In addition, nudges are often defined so as to exclude the use of incentives, whereas PTIs often do employ incentives. As we discuss in greater detail below, however, both nudges and PTIs aim to alter the choice environment so as to promote positive

behaviours and circumvent the explicit and implicit biases that often prevent individuals from exhibiting these behaviours, which it is thought they themselves would rationally endorse under the appropriate conditions. Thus, while PTIs and nudges often achieve these ends in different ways, as we explain below, discussions of the ethics of nudging provide a helpful starting point, particularly given the dearth of similar discussions regarding PTIs in particular.

2 An Exemplar PTI

As noted above, PTIs are interventions that aim to promote perspective-taking by altering the choice environment in ways that incentivize perspective-taking in order to succeed at some other explicit task. In order to assess the moral permissibility of PTIs, we will focus on one particular PTI as a case study:

> ***Town Hall.*** A city's mayor wants to improve recently fraught relations between citizens of her city. Deep and antagonistic disagreements across political lines have resulted in increased civil unrest and have discouraged tourists from visiting during peak summer months, thereby threatening a key source of the city's revenue. Upon the advice of members of the city council, she devises a plan to combat this issue by bringing citizens together under the auspices of deciding how to allocate city resources. But the mayor's real goal is to induce perspective taking amongst the citizens, given the evidence suggesting that perspective taking can help to reduce bias, promote liking, and improve relationships across various group divides. Drawing on vast evidence supporting the role of cooperative, interdependent exercises for encouraging perspective taking in schools, the exercise will involve dividing citizens into smaller diverse groups and tasking each group with the goal of identifying the most pressing city issue from a certain area (e.g. parks and recreation, education, residential and employment tax redistribution, etc.) for which resources are needed. Each member of the group will be responsible for researching and and teaching others about a particular issue, and a resolution can be reached only when a neutral, third-party city official determines that each group member has fully comprehended all of the relevant issues (by administering a brief test) and the group has reached a full consensus regarding which issue to allocate resources to addressing. Citizens must therefore cooperate and depend on one another in order to successfully complete the task at hand, a process that is greatly aided by perspective taking, including imagining how to convey information in ways that other group members will understand and how to ask questions in ways that will elicit valuable information. They are also offered a significant sum of money as compensation for reaching a full consensus, thereby strongly incentivizing full and active participation in the intervention. It is decided that the real goal of the exercise and how it works, and that fact that allocation of resources is not the real goal, will be kept hidden from participants in order to ensure that explicit and implicit biases, as well as general dislike and prejudice, do not inhibit citizens from taking part in the activity and maximally benefitting from the intervention. Real discussions about how to allocate the same city resources will be held two weeks later by city officials alone.[1]

[1] See Davis and Maitner's (2009) discussion of the benefits of cooperative-interdependent activities like the one described in *Town Hall* for inducing motivation to perspective-taking across various group divides.

In this case, individuals are encouraged to take others' perspectives in order to promote one putatively positive outcome—namely, the effective allocation of shared resources to address joint problems of mutual concern, while the real goal is to encourage perspective-taking across group divides and thereby reduce intergroup antagonism and civil unrest. Although they are not not explicitly required to do so, participants are implicitly incentivised to take others' perspectives in order to gain valuable insights into the thoughts and feelings of their fellow group mates, which in turn allows them to better convey and receive crucial information needed to pass the test for comprehension administered by the city official and reach a unanimous consensus with their group. Because the goals of promoting perspective taking across group divides and reducing intergroup antagonism and civil unrest are hidden from participants, any disinclination to forge more positive relationships or imagine others' viewpoints due to bias, prejudice, or strong dislike for outgroup members is mitigated.

Although our exemplar PTI is imagined, other real-world PTIs share important features with it. Consider, for example, the widely-used Jigsaw Classroom cooperative learning and teaching model. Developed by psychologist Eliot Aronson and implemented in Austin, TX classrooms during desegregation in the 1970s, the Jigsaw Classroom was aimed at promoting perspective taking, empathy, and mutual respect in newly desegregated and racially diverse classrooms (Aronson 2002). In order to avoid students opting out of the interventions or participating in bad faith due to prejudice towards their fellow classmates, the students were not explicitly told that a goal of the exercise is to take others' perspectives, empathize, gain mutual respect, and so on. Instead, students were placed into racially diverse groups where they were forced to cooperate and depend on one another in order to succeed at a shared academic task.[2] The Jigsaw Classroom approach thus meets the basic criteria outlined by Allport's (1954) well-known Contact Hypothesis for interventions aimed at promoting positive interactions, perspective taking, empathy, cooperation, and reduced prejudice (Pettigrew and Tropp 2005; Tropp and Saxena 2018). It continues to be widely implemented in classrooms for children and adult students alike with great success in terms of achieving its desired outcomes.[3]

3 Reasons for Doubting the Permissibility of PTIs

It is, we think, intuitively plausible that the PTI introduced in the previous section is morally permissible, particularly given the seemingly good end of promoting positive constructive interactions between politically divided citizens. However, in this section, we will call its permissibility, and that of some PTIs, such as the Jigsaw Classroom, into question by drawing attention to the ways that individuals' autonomy may be undermined by a given PTI's lack of transparency and resistibility.

Furthermore, although we do not wish to make any empirical claim regarding the frequency with which real-world PTIs are non-transparent or difficult to resist, we suspect that

[2] For instance, in order to complete a project on WWII, each group member might be tasked with researching and becoming an expert on some aspect of the material while everyone is tested on all aspects of the material. In this way, group members have to take one another's perspectives in order to determine how to most effectively convey their information such that that their fellow group mates understand, as well as ask questions to effectively elicit information that they themselves need in order to succeed at the assignment.

[3] See Perkins and Saris (2001) for a real-life case of the "Jigsaw Classroom" in a college-level statistics course. See also Read (2023) for additional background on and discussion of the Jigsaw Classroom for promoting empathy and perspective taking.

PTIs will have these features in a range of important cases. In particular, it would not be surprising to find that those implementing PTIs often choose non-transparent and difficult-to-resist interventions in cases where participants are likely to opt out of the intervention for bias or prejudice-related reasons and the implmenetor wishes to maximize the positive effects of the intervention for participants. The Jigsaw Classroom case is a paradigmatic example of this type of situation.

While the moral permissibility of PTIs that are non-transparent or difficult to resist has avoided scrutiny, concerns regarding lack of transparency and resistibility have been raised in the context of recent discussions concerning the ethics of nudging. We therefore take these discussions as a starting point for assessing the permissibility of the *Town Hall* intervention.

Consider first the transparency concern. It is often claimed that permissible nudges must be sufficiently transparent so as not to undermine individuals' autonomy by changing the choice environment in ways that are opaque and leave them no choice but to act in accordance with the nudge (Ivanković and Engelen 2019; Schmidt 2017). Yet, there is significant disagreement concerning what kind of transparency is important. On some formulations, a nudge can be transparent even if no information about the nudge is explicitly provided. For example, according to Thaler and Sunstein (2009), nudge transparency is simply a matter of the relevant nudge being susceptible to a public defense. Most formulations, however, are stronger than this. Some so-called "explicit disclosure' formulations'' take it that explicit information is needed (De Marco and Douglas 2022). For instance, Hausman and Welch (2010) argue that transparency involves the explicit provision of information about what the nudge aims to achieve and how it is expected to achieve it. Other 'easy unmaskability' characterizations of transparency maintain that either typical or particularly watchful nudgee are easily able to discover such information or 'unmask' the nudge.[4]

In the case of some PTIs, the goal of promoting perspective-taking is not explicitly disclosed to participants, nor is it easy to uncover even by the most observant individuals. In *Town Hall*, for example, the primary goal of promoting perspective taking is purposefully kept hidden so as to (1) ensure that participants are fully engaged in the activity and (2) reduce the likelihood that implicit and explicit biases they may have toward one another prevent the exercise from having its intended effects. The intervention in *Town Hall* is, by design, transparent in neither the 'explicit disclosure' nor the 'easy unmaskability' sense. Insofar as it risks undermining individuals' autonomy by failing to provide them clear opportunities for recognizing and opting out of the intervention, this lack of transparency might reasonably be considered problematic.

A second reason to doubt the permissibility of some PTIs is their lack of easy resistibility. As with transparency, although it has gone unaddressed in the case of PTIs, concerns regarding resistibility have commonly been raised in the literature on nudges and nudge-like interventions. In this literature, it is often suggested that, though nudge-like interventions could compromise autonomy by causing the targeted individual to act contrary to their prior and perhaps autonomously formed preferences, this worry can be evaded by ensuring that the intervention is easy to resist (De Marco and Douglas 2022; Thaler and Sunstein 2009). So while tweaks to the choice environment might be made to discourage

[4] Whether information must be explicitly provided, or only easily discoverable, the information might include information about the *type* of strategy to be employed and for what purposes; and/or information about how a *particular* nudge operates and what its aims are—what Bovens (2009) refers to as "type" and "token" transparency respectively. We henceforth limit our consideration to token transparency, both in relation to explicit disclosure, and easy unmaskability.

unhealthy behaviours, say, if individuals can still easily eat what they like, smoke, avoid exercise, and so on, their autonomy is preserved. On the other hand, if it becomes difficult to maintain the unhealthy behaviour, autonomy will arguably have been reduced.

There are different views of what easy resistibility should involve when it comes to nudges and nudge-like interventions.[5] On Thaler and Sunstein's (2009) classic formulation, an intervention's being easy to resist means that it is "cheap and easy to avoid" (6). On Saghai's (2013) influential and more detailed account, resistibility has three parts: the target (a) has the capacity to become aware of the pressure to get her to behave in some way; (b) is sufficiently capable of down-regulating her automatic tendency to succumb to that pressure[6]; and (c) is not in a situation that seriously undermines her ability to easily exercise these two capacities.

Many typical PTIs, including *Town Hall*, are not clearly easy to resist, at least not in Saghai's sense, which holds that the target of a nudge must be sufficiently capable of down-regulating her automatic tendency to succumb to the pressure of the nudge. In the case of PTIs it is not always clear what exactly succumbing to the pressure would consist in. For example, would succumbing to the pressure imposed by the intervention in *Town Hall* consist in performing the required tasks and, in doing so, actually taking the perspectives of other group members? Would it consist in increased perspective-taking within and beyond the confines of the intervention? Or would it consist in actually changing one's future behaviour towards members of other groups? In the interests of charity to an interlocutor who claims that PTIs are easy to resist, we assume that the primary intention of the mayor in *Town Hall* is to prevent antagonistic behaviour towards members of different social groups beyond the confines of the intervention. Thus, we assume that one succumbs to the pressure exerted by the intervention in *Town Hall* only if one subsequently takes acts less antagonistically at least in part as a causal consequence of the intervention.

Given this understanding, there are four broad possibilities for resisting the PTI in *Town Hall*: (i) opt out of the PTI altogether, (ii) participate in the intervention but avoid perspective-taking, (iii) participate in the intervention but take others' perspectives only within the confines of the intervention, and (iv) take others perspectives beyond the confines of the intervention, but continue acting just as antagonistically towards members of different social groups. In *Town Hall*, participants may not be in a position to easily resist the PTI in any of these ways. We assume that individuals are able to opt-out of the deliberative process in *Town Hall*. However, since they are unaware, and cannot easily become aware, of the PTI being employed in this case, it is doubtful that opting out of the deliberative process would count as resisting the PTI in the sense that is relevant for preserving autonomy. Participants may opt out of the intervention for various reasons (for example, a conflict with their schedule or a lack of civic engagement), but this would not help to preserve the autonomy of those who do not wish to take the perspectives of others.

Could participants resist the PTI in the second way: by participating in the intervention but avoiding perspective taking? Perhaps. But this will depend on how strong the incentive is to take others' perspectives. Suppose it is very strong. Suppose, for example, that individuals are offered a strong monetary incentive to participate in the intervention that involves perspective taking, as in the case of *Town Hall*. In that case, resisting the pressure to take others' perspectives will not be easy.

[5] For discussion, see De Marco and Douglas (2022).

[6] We assume that, to succumb to the pressure of the nudge is to act in the way that the nudger intends at least partly as a causal consequence of the nudge.

Consider next whether participants could easily resist the PTI by taking others' perspectives only within the confines of the case.

This may be difficult because, once one has seen something from another's point of view, it can be difficult to 'unsee' it. Think, for instance, of the difficulty of unseeing the effects of an optical illusion. There are also reports of individuals who, having taken others' perspectives come to appreciate defects in their past conduct towards those individuals, cannot return to their old ways of seeing and doing things.

Consider, for example, the case of Meghan Phelps-Roper for whom encounters with ideological opponents online afforded her the opportunity to appreciate the wrongness of her previous actions as part of the Westboro Baptist Church (an American Calvinist hate group). As Phelps-Roper herself describes, "Once I saw that we were not the ultimate arbiters of divine truth but flawed human beings, I couldn't justify our actions, especially our cruel practice of protesting funerals and celebrating human tragedy. And eventually, it made it impossible for me to stay."[7] Unable to 'unsee' things from the perspective of those she once targeted, Phelps-Roper radically and permanently altered her behavior: she left the church, cut ties with most of her family, and campaigns in favor of compassion and empathy across fraught ideological divides. Of course, *Town Hall* is unlikely to induce such a stark and significant change in attitudes as the one undergone by Phelps-Roper, but it could nevertheless involve changes in perspective that are difficult to 'unsee'.

Finally, consider whether participants in *Town Hall* could resist only the ultimate behavioural effects of the intervention. Suppose that the intervention succeeds in inducing participants to take the perspectives of members of different social groups to a greater degree than previously, even beyond the confines of the intervention. Could those participants nevertheless continue acting just as antagonistically towards the members of those groups? Presumably they could, but we see no reason to suppose that it would be *easy* for them to do so. As the Phelps-Roper example mentioned above suggests, changes of perspective can produce powerful changes in motivation.

In light of these considerations, there is *prima facie* reason to worry that the *Town Hall* intervention is neither transparent nor easy to resist, and thus is morally problematic. In the next two sections, however, we offer two lines of argument in support of the view that *Town Hall* is permissible even if it is indeed non-transparent and hard-to-resist (by which we mean that it is not easy to resist on Saghai's (2013) formulation, and is transparent in neither in the sense that it is explicitly disclosed nor in the sense that typical targets of the intervention can easily 'unmask' it.

4 Defending *Town Hall* I: The Appeal to Deceptive Research

Our first line of argument draws an analogy between *Town Hall* and forms of deception employed in psychological research that are often thought to be permissible. We focus our discussion on morally permissible deceptive research in psychology in particular, both because this sort of research involves interventions that bear the most similarity to *Town Hall* (as compared with, e.g., medical research), and also because there is significantly less consensus regarding morally permissible deception for other types of research.

Deception in psychological research is often thought to be morally permissible, even when it involves subjecting participants to hard-to-resist influences, such as strong negative

[7] https://www.npr.org/transcripts/560181511

emotions in response to witnessing another person's suffering (Batson et al. 2015). For example, according to the American Psychological Association's *Ethical Principles of Psychologists and Code of Conduct* (Behnke 2009), deception in psychological research is permissible provided that the following conditions are met:

a. Psychologists do not conduct a study involving deception unless they have determined that the use of deceptive techniques is justified by the study's significant prospective scientific, educational, or applied value and that effective nondeceptive alternative procedures are not feasible.
b. Psychologists do not deceive prospective participants about research that is reasonably expected to cause physical pain or severe emotional distress.
c. Psychologists explain any deception that is an integral feature of the design and conduct of an experiment to participants as early as is feasible, preferably at the conclusion of their participation, but no later than at the conclusion of the data collection, and permit participants to withdraw their data.

While the guidance provided by the APA hardly constitutes the final word when it comes to research ethics in general or psychological research ethics in particular, similar conditions have also been endorsed by others.[8]

Condition (a) is likely to be met in cases where the research is expected to generate sufficiently significant scientific or social value. Relevant to determining whether this condition has been met is the extent to which deception is required in order to protect against biassed results (Eckerd et al., 2021; Wendler and Miller 2004). For example, informing participants that they will be performing an activity with a computer may affect results in undesirable ways—for example, by generating results that are not generalizable to interactions apart from those between humans and computers. A mild form of deception, in this case to achieve more generalizable results, might involve simply replacing "computer" with "a partner" in the instructions given to participants. This may lead participants to believe that they are playing against another human, thereby generating results that generalize to human–human interactions. In this way, deception may prove crucial to addressing certain research questions.

Condition (b), concerning the extent to which subjects are subjected to "physical harm or severe emotional distress" will typically be satisfied in psychological research provided that participants are not confronted with trauma-inducing or otherwise significantly harmful stimuli. And condition (c) could plausibly be met by informing participants of the intervention's mechanisms and intended effects as soon as possible after the intervention.

Taking these putative jointly sufficient conditions for permissible deceptive research as a starting point, we suggest, first, that non-transparent and hard-to-resist PTIs *used in research* could meet all of these desiderata and thus, if the conditions are correct, be morally permissible. We then suggest that this provides some support to the view that the *Town Hall* intervention, which occurs outside a research context, would also be permissible.

There is no general reason to suppose that research involving non-transparent and hard-to-resist PTIs could not satisfy conditions (a), (b) and (c). With respect to condition (a), the 'absence of alternatives' component of this condition is likely to be satisfied for PTIs employed in research when either (i) the research is specifically examining the effects of deception (for example, whether non-transparent PTIs are more effective than transparent ones), or (ii) the

[8] See, for example, the Meta-code of Ethics developed by the European Federation of Psychologists' Association (Lindsay 2021)

research is examining psychological phenomena that can be most effectively produced through non-transparent PTIs (for example, a study on the nature of altruism towards out-group members, where that altruism can only effectively be elicited through the use of non-transparent PTIs (Zaki & Cikara, 2015)). There is also no reason to think that research involving PTIs could not generate sufficient scientific and social value so as to meet the 'valuable research' component of condition (a). Condition (b), concerning the risk of harm, will typically be satisfied provided that the PTI-involving research does not confront participants with trauma or distress-inducing stimuli. And finally, the partial consent procedures typically employed in deceptive research, which normally involve disclosure of relevant information after the deceptive intervention has been administered, could be employed in relation to research involving non-transparent and hard-to-resist PTIs, thereby satisfying condition (c). Our discussion to this point suggests that non-transparent and hard-to-resist PTIs employed in research could sometimes be permissible, at least if conditions (a)-(c) outlined above are indeed jointly sufficient for permissibility. But this in turn suggests that similar PTIs employed *outside* research, like the *Town Hall* intervention, could also sometimes be permissible. Research is, of course, a special context, and it is commonly thought that the potentially large benefits of research, and the strong safeguards to which it is subject, justify certain practices in research that would typically not be justified in other contexts (Rid and Wendler 2011). Examples include deliberately infecting people with a dangerous pathogen, exposing a person to an experimental drug, and even having tissue removed as part of a medical intervention (Kapp 2006).

However, it is doubtful that research is *unique* in the benefits it can yield and the safeguards to which it is subject. It is true that psychological research can have immense social benefits. Consider, for example, that past psychological research has unmasked the extent and costs of implicit biases (Stewart et al. 2003), discovered simple means of promoting altruistic behaviour (Darley and Batson 1973), demonstrated the effectiveness of psychological therapies (such as cognitive-behavioural therapy for anxiety disorders) (Kendall et al. 2008), and enabled the development of highly impactful public health interventions (such as efforts to promote social distancing in response to the COVID-19 pandemic) (Bonell et al. 2020). However, though it is typically *possible* that any particular psychological study will have immense benefits, the *ex ante* likelihood that it will do so is normally very small. Thus, at least in typical cases, the expected net social benefits of an individual psychological experiment are, though significant, rather modest. One can imagine high-stakes contexts in which the use of PTIs outside of research might have benefits that are at least comparable to the typical benefits of research. One example would be the use of PTIs as part of 'citizens assemblies' or other deliberative-democratic procedures intended to decide on important matters of policy, such as abortion law (Suiter et al. 2016). Another would be the use of PTIs in jury deliberations concerning conviction for serious crimes. If it were shown that non-transparent and hard-to-resist PTIs significantly improved the quality of deliberation in these contexts, the case in favour of employing them would, given the importance of the decisions made in these contexts, be strong. Given the significant costs that intergroup antagonism can have (Sinnott-Armstrong 2018), we propose that the *Town Hall* intervention would, if significantly more effective than alternative interventions, likely have expected benefits that would be at least comparable to those of a typical psychological experiment involving deceptive techniques.

Moreover, given that *Town Hall* occurs in an unusual and well-defined context, it is plausible that this PTI could be subjected to safeguards comparable to those employed in research. We thus find it hard to see how one could maintain the permissibility of deceptive psychological research employing hard-to-resist techniques, while denying that the *Town Hall* PTI could be permissible.

Of course, we may well have missed some important disanalogy between the *Town Hall* PTI and morally permissible deceptive psychological research. In the next section, we therefore consider a second line of defense for the view that the *Town Hall* PTIs may indeed be permissible even if non-transparent and hard to resist.

5 Defending *Town Hall* II: Drawing Attention to Reasons

A second defence of the *Town Hall* PTI appeals to the way in which it operates by making salient the consequences of participants' actions on others, thereby drawing attention to normative reasons to think or act in a particular way. For example, in inducing participants to take the perspectives of members of other social groups, the *Town Hall* PTI plausibly makes salient to them the reasons they have not to act antagonistically towards members of those groups (for example, because doing so is likely to occasion feelings of exclusion or distress). In making salient one's normative reasons (henceforth just reasons), the *Town Hall* PTI is akin to a widely employed and relatively uncontroversial type of nudge: informational nudges.

Informational nudges operate by explicitly providing information or reminders in order to change the salience of various options. For instance, opportunity-cost reminder nudges provide information that increases the salience of what will be foregone if money (or time) is spent at one point rather than another (e.g., today rather than tomorrow) (Thunström et al. 2018). Reminders to schedule follow-up dental appointments that explicitly mention the low cost and long-term benefits of preventative dental care are an example of this (Altmann and Traxler 2014). Descriptive social-norm nudges, by contrast, describe others' behaviour in order to increase the salience of a social norm, and thus enhance compliance with that norm (Bicchieri and Dimant 2019). An example of this is issuing a reminder such as "nine out of ten people pay their taxes on time," in order to encourage timely tax payments (Hallsworth et al. 2017).

Informational nudges do not typically present information that is explicitly normative (such as 'you ought to pay your taxes on time'), however they do typically operate by drawing the attention of nudgees to considerations that plausibly indicate, and are likely to be seen by nudgees as indicating, the presence of a reason. Thus, such nudges typically draw the nudgee's attention to their putative reasons. Consider social-norm nudges. These present descriptive information about the behaviour of one's peers which plausibly indicates, and which nudgees are likely to see as indicating, the presence of normative reasons to perform the desired action (i.e., to conform to the social norm). In this way, they draw the nudgees' attention to their putative reasons for action. Social-norm nudges can thus be seen as an implicit form of rational persuasion.[9]

Similar thoughts apply to most other informational nudges. Think, for instance, of campaigns to promote vaccination by describing the risks to others posed by refusing to vaccinate oneself (Shilo et al. 2021), discourage poor driving behavior by highlighting the harms that can result from vehicle collisions (Cismaru and Nimegeers 2017), and discourage smoking in the vicinity of one's children by depicting children suffering from serious lung disease (Clayton et al. 2020). In these cases, attention is drawn to the nudgee's putative reasons for action: receiving a vaccine will be safer for one's elderly and newborn loved ones, driving safely will protect the driver, passengers, other drivers, and pedestrians, and smoking near children exposes them to the many harms of second-hand smoke.

[9] For a defence of the more general view that nudges typically operate by implicitly giving reasons, see Levy (2019).

Like informational nudges, the *Town Hall* PTI plausibly promotes the desired (less antagonistic) behavior by implicitly drawing participants' attention to their putative reasons for action. Yet, as we described above, it is also plausible that this intervention is neither transparent nor easy to resist. Does this render it impermissible? Reflection on informational nudges suggests that it does not. For, as we will now argue, informational nudges (or informational nudge-like interventions) sometimes *also* fail to meet these criteria while remaining, intuitively, morally permissible.[10]

Consider the practice of discouraging speeding through the use of road markings that draw drivers' attention to the fact that someone is driving quickly. Like the informational nudges described above, this intervention makes salient the fact that one is driving very quickly, which plausibly indicates the presence of a reason to slow down, and which the target is likely to see as such. It will thus, in typical cases at least, draw the nudgee's attention to the putative reason. Yet we can suppose that the intervention is non-transparent in both the explicit disclosure and easy unmaskability senses; drivers are not explicitly told why these particular road markings have been chosen and nor is it easy for drivers to infer the intent behind the markings (we might suppose that they look very similar to ordinary road markings). It is also not clear that the intervention will always or even typically be easy to resist. Perhaps many drivers, when confronted in a salient way with the speed at which they are driving, experience a strong urge to slow down. Nevertheless, most will, we suspect, regard this intervention as permissible. And we think there is a good explanation for *why* it is permissible: it operates by drawing its target's attention to their reasons to do the desired action.

Something similar may be said of an intervention that aims to draw potential offenders' attention to their reasons for not committing a crime by installing better lighting in a parking lot or other public place. In such a case, we suspect that most will find an intervention such as this to be permissible despite being non-transparent and difficult to resist. After all, there is no explicit disclosure of the reason for the better lighting, it is likely difficult to unmask the reason for installing the improved lighting simply from observing it, and potential offenders will likely find it difficult to avoid being affected by the fact that there is better lighting in the place.

If this explanation is correct, we have support for the view that PTIs—even when neither transparent nor easy to resist—will also be morally permissible when, as is plausibly the case for the *Town Hall* intervention, the PTI operates by drawing participants' attention to their reasons for action.[11]

6 An Objection and Reply

At least one important objection could be raised against the defenses of the *Town Hall* PTI offered above: it arguably presupposes and imposes a particular conception of the good, thereby failing to satisfy liberal neutrality, understood as a requirement to

[10] The parenthetic qualification is meant to accommodate the view that, if an intervention is not easy to resist, it is not a nudge.

[11] Though it is not our intention to offer a contribution to the debate on nudging here, we believe that intuitive responses to interventions such as the road marking intervention may also have interesting implications for the moral permissibility of nudges, for they suggest, contrary to claims sometimes made in that debate, that nudges or nudge-like interventions can sometimes be permissible even when non-transparent and hard-to-resist.

accommodate rival (reasonable) conceptions of the good. The PTI in *Town Hall* aims at promoting the putatively good outcome of reduced inter-group antagonism and civil unrest. There may, however, be reasonable conceptions of the good on which this is not a valuable outcome. Thus, one may worry that, despite promoting purportedly good ends, PTIs fail to respect reasonable disagreement concerning the nature of the (moral or prudential) good.

Against this objection, we maintain that some conceptions of the good, including those due to racist or otherwise bigoted ideologies, are unreasonable, so that any failure on the part of PTIs to remain neutral with respect to them is in fact permissible. According to Rawls (1999), reasonableness requires, among other things, a basic readiness to cooperate with others on terms that are both fair and acceptable to other reasonable persons. Bigoted conceptions of the good are unlikely to meet this condition. So PTIs need not be required to respect them.[12] Thus, insofar as the *Town Hall* PTI is intended to overcome barriers to perspective-taking that are due to acceptance of a bigoted conception of the good, it will not violate liberal neutrality.

Moreover, it is consistent with the aims of liberal neutrality to promote a particular conception of the good, even at the expense of other reasonable conceptions, if one does so by drawing attention to reasons that speak in favor of it. In fact, because even reasonable persons will sometimes disagree about their conceptions of the good and the moral values that underlie them—what Rawls (2005) refers to as the possibility of *reasonable pluralism*—negotiating these differences for the sake of engaging in joint deliberation or cooperation toward shared goals may sometimes require that one or more parties give reasons in support of some conception of the good. A familiar example of this is the graphic health warnings on cigarette packages, which might be said to draw attention to the self- and other-regarding reasons for acting in ways that accord with a conception of the good on which one's own and others' health is valuable and worth protecting.[13]

Similarly, PTIs frequently draw attention to reasons for acting in accord with a conception of the good that places disvalue on causing gratuitous harm to others or inhibiting their ability to fully participate in public life. The PTI in *Town Hall* draws individuals' attention to reasons to be less antagonistic towards members of different social groups. These reasons will likely include facts about the likely negative effects of antagonistic treatment on members of different groups–effects that one comes to appreciate by taking the perspective of the members of those groups. Reason-giving of this sort is permissible, even in the face of disagreement about the underlying conceptions of the good.[14]

[12] Our claim is in keeping with Tillson's (2019) view that some, but not all, forms of "formative influence" are impermissible, to the extent that they limit individuals' awareness of the range of answers to questions about "what we have most reason to value and how we ought to treat one another" (2). See also Clayton's (2006) argument against the unbounded rights of parents to impose values and conceptions of the good on their children when doing so violates the demands of political legitimacy to exercise "authority in accordance with public reason, in a way that is capable of acceptance by free and equal persons" (94).

[13] See examples in both Europe (https://ec.europa.eu/health/tobacco/law/pictorial_en) and the U.S. (https://www.fda.gov/tobacco-products/labeling-and-warning-statements-tobacco-products/fda-proposes-new-health-warnings-cigarette-packs-and-ads).

[14] In addition to drawing attention to pre-existing reasons, the *Town Hall* PTI also plausibly creates prudential reasons, in the form of incentives, to engage in the tasks that are intended to result in the perspective-taking. However, since the use of incentives is not normally thought to be morally problematic, except in certain special cases, we do not consider this aspect of PTIs further. We thank an anonymous reviewer for pressing this point.

7 Concluding Remarks

This article has considered the moral permissibility of PTIs: interventions that promote taking the perspectives of others by incentivising or otherwise inducing individuals to take the perspectives of others. Despite advancing morally valuable ends, PTIs can be morally questionable because they are neither transparent nor easy to resist. Nevertheless, we defended a plausibly non-transparent and hard-to-resist PTIs—the *Town Hall* intervention—by drawing on analogies between it and both deceptive research and certain morally permissible nudges or nudge-like interventions.

Finally, we replied to one important objection to our defences—namely, that the *Town Hall* PTI violates liberal neutrality. In response to this objection, we maintained that (i) some conceptions of the good are unreasonable such that any failure on the part of PTIs to remain neutral with respect to them is in fact permissible; and (ii) it is consistent with the aims of liberal neutrality to promote a particular conception of the good if one does so by giving reasons to adopt or act in a way that accords with it, as is plausibly the case with the *Town Hall* PTIs.

We have focussed throughout on a particular PTI: the one employed in *Town Hall*. However, our arguments will of course extend to many other PTIs as well, including others that are non-transparent and hard to resist. We leave the precise specification of the range of PTIs that are morally permissible as a task for future research. However, our analysis suggests several features that will be relevant to determining their permissibility. Perhaps most important among these are (i) whether the PTI has substantial benefits that could not be realised via other means, (ii) whether the PTI is subject to safeguards akin to those employed for deceptive research, (iii) whether the there a reasonable conceptions of the good on which the intended goal of the PTI is disvaluable, and (iv) whether the PTI operates by drawing participants' attention to reasons.[15]

Author's Contribution HR and TD jointly conceived the argument of the paper. HR wrote an initial draft, and HR and TD then each performed multiple rounds of substantial edits in the light of joint discussions of the argument.

Funding TD received funding from the European Research Council [grant number 819757] and the Uehiro Foundation on Ethics and Education.

Data Availability N/A.

Declarations

Ethical Approval N/A.

Informed Consent N/A.

Competing Interests TD has received funding, for work unrelated to this paper, from Merck KGaA, Darmstadt.

Statement Regarding Research Involving Human Participants and/or Animals N/A.

[15] We are very grateful for helpful feedback on earlier drafts of this paper from Gabriel De Marco, Maximilian Kiener, John Tillson, and participants of the Workshop on the Ethics of Influence at Oxford University.

References

Allport GW (1954) The nature of prejudice. Perseus Books, Cambridge, MA
Altmann S, Traxler C (2014) Nudges at the dentist. Eur Econ Rev 72:19–38
Arno A, Thomas S (2016) The efficacy of nudge theory strategies in influencing adult dietary behaviour: A systematic review and meta-analysis. BMC Public Health 16(1):676
Aronson E (2002) Building empathy, compassion, and achievement in the jigsaw classroom. In Improving academic achievement (pp. 209–225). Elsevier
Batson CD, Early S, Salvarani G (1997) Perspective taking: Imagining how another feels versus imaging how you would feel. Pers Soc Psychol Bull 23(7):751–758
Batson CD (2012) Two Forms of Perspective Taking: Imagining How Another Feels and Imagining How You Would Feel in Markman, K. D., Klein, W. M., & Suhr, J. A. (Eds.), Handbook of imagination and mental simulation. Psychology Press
Batson CD, Lishner DA, Stocks EL (2015) The empathy—Altruism hypothesis. In Schroeder, D.A., and Graziano, W.G., The Oxford Handbook of Prosocial Behavior. Oxford: Oxford University Press
Behnke S (2009) Ethics rounds: Reading the Ethics Code more deeply. Monitor Psychol 40(4) https://www.apa.org/monitor/2009/04/ethics
Bicchieri C, Dimant E (2019) Nudging with care: The risks and benefits of social information. Publ Choice 1–22
Bloch KE (2021) Virtual reality: Prospective catalyst for restorative justice. Am Crim l Rev 58:285
Bonell C, Michie S, Reicher S, West R, Bear L, Yardley L, ... Rubin GJ (2020) Harnessing behavioural science in public health campaigns to maintain 'social distancing' in response to the COVID-19 pandemic: key principles. J Epidemiol Community Health 74(8) 617–619
Bovens L (2009) The Ethics of Nudge. In: Grüne-Yanoff T, Hansson S (eds) Preference change approaches from philosophy. Springer, Economics and psychology, pp 207–219
Bridgeman DL (1981) Enhanced role taking through cooperative interdependence: A field study. Child Dev 1231–1238
Broockman D, Kalla J (2016) Durably reducing transphobia: A field experiment on door-to-door canvassing. Science 352(6282):220–224
Cismaru M, Nimegeers K (2017) "Keep your eyes up, don't text and drive": a review of anti-texting while driving Campaigns' recommendations. Int Rev Publ Nonprofit Market 14(1):113–135
Clayton M (2006) Justice and legitimacy in upbringing. Oxford University Press on Demand, Oxford
Clayton RB, Keene JR, Leshner G, Lang A, Bailey RL (2020) Smoking status matters: A direct comparison of smokers' and nonsmokers' psychophysiological and self-report responses to secondhand smoke anti-tobacco PSAs. Health Commun 35(8):925–934
Coplan A (2011) Will the real empathy please stand up? A case for a narrow conceptualization. South J Philos 49:40–65
Darley JM, Batson CD (1973) " From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. J Pers Soc Psychol 27(1):100
Davis MH, Maitner AT (2009) Perspective taking and intergroup helping. The psychology of prosocial behavior: Group processes, intergroup relations, and helping, 173–190
De Marco G, Douglas T (2022) Nudge Transparency Is Not Required for Nudge Resistibility. Ergo., online first.
Eckerd S, DuHadway S, Bendoly E, Carter CR, Kaufmann L (2021) On making experimental design choices: discussions on the use and challenges of demand effects, incentives, deception, samples, and vignettes. J Oper Manag 67(2):261–275
Galinsky AD, Moskowitz GB (2000) Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. J Pers Soc Psychol 78(4):708

Gehlbach H, Marietta G, King AM, Karutz C, Bailenson JN, Dede C (2015) Many ways to walk a mile in another's moccasins: Type of social perspective taking and its effect on negotiation outcomes. Comput Hum Behav 52:523–532
Greenberg MT, Domitrovich CE, Weissberg RP, Durlak JA (2017) Social and emotional learning as a public health approach to education. The future of children, 13–32
Gutsell JN, Simon JC, Jiang Y (2020) Perspective taking reduces group biases in sensorimotor resonance. Cortex 131:42–53
Hallsworth M, List JA, Metcalfe RD, Vlaev I (2017) The behavioralist as tax collector: using natural field experiments to enhance tax compliance. J Public Econ 148:14–31
Hausman DM, Welch B (2010) Debate: to nudge or not to nudge. J Polit Philos 18(1):123–136
Ivanković V, Engelen B (2019) Nudging, transparency, and watchfulness. Soc Theory Practi 43–73
Kapp MB (2006) Ethical and legal issues in research involving human subjects: do you want a piece of me? J Clin Pathol 59(4):335–339
Kendall PC, Hudson JL, Gosch E, Flannery-Schroeder E, Suveg C (2008) Cognitive-behavioral therapy for anxiety disordered youth: a randomized clinical trial evaluating child and family modalities. J Consult Clin Psychol 76(2):282
Levy N (2019) Nudge, nudge, wink, wink: Nudging is giving reasons. Ergo 6
Lindsay G (2021) The Development of the European Federation of Psychologists' Associations' Meta-Code of Ethics. In Handbook of international psychology ethics (pp. 174–187). Routledge
Muradova L (2021) Seeing the other side? Perspective-taking and reflective political judgements in interpersonal deliberation. Political Stud 69(3):644–664
Myers MW, Laurent SM, Hodges SD (2014) Perspective taking instructions and self-other overlap: different motives for helping. Motiv Emot 38:224–234
Perkins DV, Saris RN (2001) A" jigsaw classroom" technique for undergraduate statistics courses. Teach Psychol 28(2):111–113
Pettigrew TF, Tropp LR (2005) Allport's intergroup contact hypothesis: Its history and influence. On the nature of prejudice: Fifty years after Allport 262–277
Rawls J (1999) A Theory of Justice. Harvard University Press. Revised edition, Cambridge, MA
Rawls J (2005) Political Liberalism, 2nd edn. Columbia University Press, New York
Read H (2023) Institutionalized empathy. J Moral Educ 52(2):224–243
Rid A, Wendler D (2011) A framework for risk-benefit evaluations in biomedical research. Kennedy Inst Ethics J 21(2):141–179
Rithalia A, McDaid C, Suekarran S, Myers L, Sowden A (2009) Impact of presumed consent for organ donation on donation rates: A systematic review. BMJ 338(January):a3162
Saghai Y (2013) Salvaging the concept of nudge. J Med Ethics 39(8):487–493
Schmidt AT (2017) The power to nudge. Am Political Sci Rev 111(2):404–417
Shilo S, Rossman H, Segal E (2021) Signals of hope: gauging the impact of a rapid national vaccination campaign. Nat Rev Immunol 21(4):198–199
Simonovits G, Kezdi G, Kardos P (2018) Seeing the world through the other's eye: an online intervention reducing ethnic prejudice. Am Polit Sci Rev 112(1):186–193
Sinnott-Armstrong W (2018) Think again: How to reason and argue. Oxford University Press, Oxford
Stewart TL, Laduke JR, Bracht C, Sweet BA, Gamarel KE (2003) Do the "eyes" have it? A program evaluation of Jane Elliott's "Blue-Eyes/Brown-Eyes" diversity training exercise 1. J Appl Soc Psychol 33(9):1898–1921
Suiter J, Farrell DM, O'Malley E (2016) When do deliberative citizens change their opinions? Evidence from the Irish Citizens' Assembly. Int Polit Sci Rev 37(2):198–212
Thaler RH, Sunstein C (2009) Nudge. Yale University Press, New Haven
Thunström L, Gilbert B, Ritten CJ (2018) Nudges that hurt those already hurting–distributional and unintended effects of salience nudges. J Econ Behav Organ 153:267–28
Tillson J (2019) Children, religion and the ethics of influence. Bloomsbury Publishing, London
Todd AR, Galinsky AD (2014) Perspective-taking as a strategy for improving intergroup relations: Evidence, mechanisms, and qualifications. Soc Pers Psychol Compass 8(7):374–387
Tropp LR, Saxena S (2018) Re-Weaving the Social Fabric through Integrated Schools: How Intergroup Contact Prepares Youth to Thrive in a Multiracial Society. Research Brief No. 13. Natl Coalition School Divers
Walker I, Crogan M (1998) Academic performance, prejudice, and the jigsaw classroom: New pieces to the puzzle. J Commun Appl Soc Psychol 8(6):381–393
Weissberg RP (2019) Promoting the social and emotional learning of millions of school children. Perspect Psychol Sci 14(1):65–69
Wendler D, Miller FG (2004) Deception in the pursuit of science. Arch Intern Med 164(6):597–600

Zaki J, Cikara M (2015) Addressing empathic failures. Curr Dir Psychol Sci 24(6):471–476