The Philosophical Quarterly Vol. 00, No. 0 *2024*
ISSN 0031-8094 https://doi.org/10.1093/pq/pqae033

Diabolical devil's advocates and the weaponization of illocutionary force

Giulia Terzian and María Inés Corbalán

ArgLab-IFILNOVA, Nova Universidade de Lisboa, Portugal

A standing presumption in the literature is that devil's advocacy is an inherently beneficial argumentative move; and that those who take on this role in conversation are paradigms of argumentative virtue. Outside academic circles, however, devil's advocacy has acquired something of a notorious reputation: real-world conversations are rife with self-proclaimed devil's advocates who are anything but virtuous. Motivated by this observation, in this paper we offer the first in-depth exploration of non-ideal devil's advocacy. We draw on recent analyses of two better known discursive practices—mansplaining and trolling—to illuminate some of the signature traits of vicious devil's advocacy. Building on this comparative examination, we show that all three practices trade on a manipulation of illocutionary force; and we evaluate their respective options for securing plausible deniability.

Keywords: public discourse; devil's advocacy; mansplaining; trolling; conversational ethics; plausible deniability.

I. Introduction

This paper examines the practice of *giving voice to disagreements via devil's advocacy*. In so doing, it contributes to a burgeoning conversation, within socially minded philosophy, over the normativity of giving voice to disagreements in the public sphere (e.g. Johnson 2018c). In the literature, devil's advocacy is standardly construed as an inherently virtuous discursive move; and it is tacitly presumed to be deployed by virtuous arguers. Much like any discursive practice, however, devil's advocacy is liable to be abused. The present paper offers the first detailed exploration of this theoretical possibility. The following

Corresponding author: Giulia Terzian (giuliaterzian@fcsh.unl.pt)

passage, hereafter labelled Affirmative action for ease of reference, provides an initial intuitive illustration of the phenomenon we have in mind—what may be glossed, at a first approximation, as an inappropriate use of devil's advocacy:

> The man had waited in line for 10 minutes so he could tell me that his son had been denied admission to my law school alma mater because, unlike me, he was white and so couldn't benefit from affirmative action. [...] I recounted this infuriating story to a white friend of mine [...]. Instead of rolling his eyes along with me, my friend forced me to debate him—on behalf of the man from the panel [and] the devil—on whether maybe I really hadn't deserved admission to my law school.[1]

Something has gone seriously awry in the reported exchange. But what, exactly? One might be tempted to dismiss it as a mere glitch, perhaps reflecting the speaker's misunderstanding of the point, or purpose, of devil's advocacy. We find this unconvincing; observations such as the following, moreover, suggest that the significance of Affirmative action stretches far beyond the anecdotal:

> [C]ross-racial conversations about race have become [...] more common, and thankfully so. Unfortunately, this has invited a dangerous tendency for white people to engage in these discussions with people of color by summoning the devil himself and treating racism as a political disagreement around which two opposing viewpoints can reasonably form.[2]

Scenarios such as Affirmative action bring to salience a question that has not, to our knowledge, been explored: under what conversational circumstances is it *permissible*, or normatively appropriate, to play devil's advocate? Addressing this question, in turn, requires clarifying what exactly it is that devil's advocates *do*—in order to understand what they (can) do *wrong*. We take steps towards filling these explanatory gaps, showing that the default virtuous connotations of devil's advocacy lend themselves to being exploited, at the hands of bad faith arguers, in ways that negatively affect the deontic profiles of their conversational partners.

We proceed as follows. Section II presents two faces of devil's advocacy: the 'angelical' profile celebrated in scholarly analyses, and its non-ideal, 'diabolical' counterpart found in scenarios such as Affirmative action. Section III examines two better understood discursive behaviours—mansplaining and trolling—that are seen to resemble vicious devil's advocacy in key performative respects. This comparative exercise helps bring into sharper relief some of the explanatory limitations of ideal analyses of devil's advocacy, examined

[1] https://slate.com/news-and-politics/2017/10/playing-devils-advocate-in-conversations-about-race-is-dangerous-and-counterproductive.html All webpages accessed in June 2023.

[2] https://slate.com/news-and-politics/2017/10/playing-devils-advocate-in-conversations-about-race-is-dangerous-and-counterproductive.html.

through a speech act-theoretic lens in Section IV. Among other things, ideal analyses appear to regard the manoeuvre's characteristic preamble as performatively inert; in Section V, we argue that the power of the preamble should not be underestimated. In particular, we suggest that it is functionally recognizable as a uniquely powerful force figleaf, which enables diabolical devil's advocates to deny—not just plausibly, but near unassailably—any wrongdoing. Section VI briefly contextualizes our main findings within a growing debate on the normativity of voicing objections, and identifies leads for future research.

II. Multiple profiles

II.1 Ideal devil's advocacy

Devil's advocacy is an established means of manufacturing disagreement, familiar from our daily conversational lives. Its deployment is typically announced or flagged by highly recognizable preambles (sometimes added *post hoc*): *Let me play devil's advocate here...*, *Let's just say, for the sake of argument...*, *I was just playing devil's advocate there*, and so on. In its textbook form, moreover, devil's advocacy has well-known positive connotations. It received an illustrious endorsement from Mill, as a vital bulwark against dogmatism:

> [The truth] is [n]ever really known but to those who have attended equally and impartially to both sides and endeavored to see the reasons of both in the strongest light. So essential is this discipline to a real understanding of moral and human subjects that, if opponents of all-important truths do not exist, it is indispensable to imagine them and supply them with the strongest arguments which the most skilful devil's advocate can conjure up. (Mill 1859, 35–6)

Following Mill, a number of scholars have remarked upon the value of devil's advocacy in furthering desirable epistemic outcomes and safeguarding against cognitive and argumentative pitfalls. Emphasizing the former quality, for instance, Johnson writes:

> Paradigmatically, an agent who plays devil's advocate announces her intention to defend a position she doesn't hold, and then defends that position in order to make progress on the issue at hand. [...] Paradigmatically, the term devil's advocate describes someone who, given a certain point of view, takes a position she does not necessarily agree with for the sake of debate or to explore the thought further. (Johnson 2018a, 97, 99)

The 'angelical' virtues of devil's advocacy are extolled even more explicitly by Stevens and Cohen:

> We want an arguer who *opposes* us to *help* us: an *advocatus diaboli*, a Devil's Advocate. A devil's advocate is not merely a useful interlocutor: [...] she is the ideal other who embodies what is best and most important about argumentation. She is the opponent we need because her overall goal is to enhance the prospects of successful argumentation,

> that is, getting it right [...]. [Devil's advocates] help us transcend our limits by criticizing our argument in order to strengthen it, not to defeat it. (Stevens and Cohen 2019, 170)

Yet others have emphasized the benefits, both substantive and procedural, of devil's advocacy for our collective epistemic-deliberative practices. Thus, 'being *overtly* and *openly* uncooperative' during a group deliberation, as when one plays devil's advocate, may qualify as a virtue—since it will encourage discussants to thoroughly scrutinize their own commitments (Aikin and Clanton 2010, 419), thus allowing them to be more confident in the outcome of their deliberation (Beatty and Moore 2010).

Whether the conversational setting is one of group deliberation, a private two-person dialogue, or anything in between, the literature is clear on the fact that devil's advocacy 'find[s] a natural home in a cooperative, joint problem-solving discussion' (Jacobs 1989, 353); beyond this, no further specifications as to the boundaries of proper or acceptable uses of the manoeuvre are to be found. Thus, while we may glean that not all conversational contexts are (equally) suitable environments for playing devil's advocate, the standing presumption seems to be that speakers can be counted upon to make such determinations competently, and to uphold standards of cooperativeness and good faith, *even as* they move to inject disagreement into the conversational fray. Moreover, it is reasonable to think that the success of this balancing act depends in no small part on the *preamble*. This is perhaps most clearly seen from a politeness theory perspective: given that disagreements during social interactions are known to be face-threatening acts, we may naturally see the preamble as a recourse for introducing disagreement while minimising the corresponding threat to the hearer's positive face (Brown and Levinson 1987).[3]

Overall, the profile that emerges from the theoretical literature is unequivocally positive: devil's advocacy is uniquely placed to promote desirable epistemic, rhetorical, and argumentative goals, on both an individual and a collective level. In particular, devil's advocacy can be beneficial *to the arguer(s)*, by prompting them to critically reflect on the epistemic standing of their beliefs, helping them spot and repair weaknesses in their arguments, or equip them to face dialectical opponents. And it can be beneficial *to the argument(s)*, either in and of themselves (by uncovering hidden weaknesses), or as touchstones of the state of a particular debate (by broadening the space of conceivable options), or as benchmarks of deliberative quality (by safeguarding against groupthink

[3]See also (Goffman 1959; Terkourafi 2015). In the conversational analysis literature, disagreements are seen as the less-favoured component of a vast majority of adjacency pairs (e.g. Sacks and Schegloff 1979). See also (Aikin and Clanton 2010, 419): engaging with devil's advocates allows arguers to become 'more practiced in the art of the adversarial argumentative dialogue, but in a way that nevertheless preserves the broader cooperative background'. See also (Searle 1976) on announcements.

and intellectual stagnation).[4] Last but not least, devil's advocates are implicitly presumed to choose their moments well, and to act competently and in good faith.

II.2 Devil's advocacy in the wild

Extant analyses of devil's advocacy are heavily idealized. On the one hand, they portray devil's advocates as presumptively virtuous arguers, as we've just seen. On the other, they significantly under-describe the contextual richness of real-world exchanges. Among other things, they abstract away from: the situational features of the conversational participants, including their respective social positions and their epistemic standing with respect to the topic of the exchange; the common ground against which the conversational participants, and the exchange itself, may be measured[5]; the kind, or genre, of conversation that is taking place, and attendant rules of play.[6] As a result, extant accounts are explanatorily limited with respect to the target non-ideal phenomenon. They can handle serendipitous cases, in which presumptions of all-round virtuosity are met and no other defeating considerations stand in the way of devil's advocates making their move. No doubt, serendipitous (or near enough) conditions are (liable to be) instantiated in real-world exchanges, be it in formal settings such as classrooms, boardrooms, and debate clubs, or in casual conversation among friends, social media acquaintances, etc. Owing to their in-built context-insensitivity, however, extant accounts stumble when faced with less-than serendipitous scenarios, featuring less-than ideal devil's advocates.

Our first task, then, is to bring more contextual details into relief. For this purpose, we draw on a selection of narrative excerpts which, like Affirmative action, serve simultaneously as illustrations of the target phenomenon, and as motivation for our investigation. We may begin by observing that vicious uses of devil's advocacy recur in conversations about race- and gender-based discrimination, typically enforcing an allocation of dialectical roles that track the very same social divides made salient by those conversational topics. We saw this in Affirmative action, and we see it also in Wage gap:

> You post that article about the wage gap on Facebook, and all of a sudden, all of these cis, white, straight dudes come out of the woodwork to remind you that the statistics are faulty, that women take more time off of work, that women just don't like STEM fields—all under the guise of "playing devil's advocate".[7]

[4] It is worth noting that devil's advocacy has also been the object of a smattering of empirical studies, whose findings are largely consistent with these profile sketches. See for instance (Nemeth *et al.* 2001a,b; Duran and Fusaroli 2017; Brohinsky *et al.* 2022).

[5] For example, (Stalnaker 2002; Camp 2018).

[6] For example, (Lewis 1979; McGowan 2019).

[7] https://everydayfeminism.com/2015/09/playing-devils-advocate/.

The target scenarios also exhibit a certain imperviousness to the rules conventionally governing specific conversation genres, associated expectations of uptake and turn-taking, and particularized score rules in play. As a result, the relevant contributions are predictably disruptive, bringing on-topic conversations to a standstill and either deterring or preventing intended participants from taking their turn. We'll have more to say about each of these points in later sections. Intuitively, at least, they are exemplified in Street harassment:

> I was asked to come on [a radio show] to talk about street harassment. [...] After being asked a few broad-sweeping questions that repeatedly prompted me to address the oft-claimed defence that street harassment is 'just a compliment' [...], the host specifically asked for other tattooed women to call in and discuss their experiences with their body art and street harassment. So I really wasn't surprised when the first call answered was from a dude. [...] His argument was that not every situation can be tied back to structural oppression, and that my point about how so-called 'harmless compliments' are actually indicative of just how much women's bodies are not respected in public spaces was absurd. [...] He just wanted to [...] give us 'the other side of the story', since my stance was 'one-sided' and 'slanted'. He just wanted to 'intelligently, rationally debate' this topic [...] under the guise of 'playing devil's advocate'—as if [I had] never heard these arguments before.[8]

Notice the very last quoted remark: a widely observed characteristic of vicious devil's advocacy is the tireless reiteration of considerations that do not, by any stretch of the imagination, qualify as intellectually stimulating, outside-the-box thinking, or otherwise beneficial. Worse, they are liable to constitute an additional cognitive-epistemic burden to be carried by the recipients:

> Some might challenge that I am shutting myself off to new ideas and censoring important opportunities for growth. But these ideas you are forcing me to consider are not new. They stem from centuries of inequality and your desperate desire to keep them relevant is based in the fact that you benefit from their existence. Let it go. You did NOT come up with these racist, misogynistic theories. We've heard them before and we are f*cking tired of being asked to consider them, just one. more. time.[9]

Importantly, as noted here, the objections being reiterated under the guise of devil's advocacy are not merely unimaginative: they coincide with still-dominant narratives in race- and gender-related discourse. This speaks directly—and damningly—to the overall epistemic quality of the proffered contributions. For, a paradigmatic presupposition of devil's advocacy is that the disputant(s) may have unduly neglected a particular hypothesis (and the possible reasons supporting it). However, there is, of course, no plausible sense in which the dominant standpoints, in race- and gender-related discourse,

[8] https://everydayfeminism.com/2015/09/playing-devils-advocate/.

[9] https://feministing.com/2014/05/30/an-open-letter-to-privileged-people-who-play-devils-advocate/.

have ever been relegated to the margins. Far from enhancing deliberative quality, the observed contributions merely reinforce the *status quo*.

In light of the foregoing, it is hardly surprising that several observers have highlighted the characteristically adversarial rather than cooperative attitude of self-proclaimed devil's advocates:

> It is especially harmful [...] when a man plays the Devil's Advocate to a woman trying to discuss feminist issues; it becomes another way to silence women and disregard our experiences. Disagreeing with someone who is promoting equal rights just 'for the sake of argument' ultimately trivialises the oppression that marginalised people face daily.[10]

> ...if you're playing devil's advocate in order to try and help someone else, find out if that person actually wants or needs your help. Unsolicited advice is frankly annoying in almost any case, but especially when it involves a long, drawn-out debate with someone you believe to be in need of convincing.[11]

Extant theoretical analyses, focusing as they do on the ideal paradigm of devil's advocacy, are (unsurprisingly) silent about almost every one of these contextual features. As a result, they are ill-equipped to make sense of what one commentator has suggestively—and rather aptly—described as the 'weaponization' of devil's advocacy. We think this phenomenon deserves attention: it is politically and theoretically important to understand exactly what goes wrong in these exchanges, as part of the ongoing collective project of understanding how speech can harm, and how speech agency may be subverted. As part of this explanatory endeavour, we seek to substantiate two related intuitions: on the one hand, that the target cases cannot be dismissed as mistakes borne of linguistic incompetence, and instead exemplify (at best cavalier, at worst bad faith) vicious discursive behaviour; on the other, that the main problem in these cases lies in what diabolical devil's advocates *do*, and not (merely) what they say. On this count, vicious devil's advocacy resembles a couple of better understood discursive phenomena, as we'll now see.

III. A devil's advocate, a mansplainer and a troll walk into a bar

We briefly present the main findings of two analyses of mansplaining and trolling, respectively due to Johnson (2020) and Connolly (2022). We then indicate how these findings shed some light on real-world abuses of devil's advocacy.

[10] https://ashamedmagazine.co.uk/opeds/why-do-cis-men-love-to-play-devils-advocatenbsp.

[11] https://the-orbit.net/brutereason/2013/08/10/how-to-be-a-responsible-devils-advocate/.

III.1 Mansplaining

Johnson (2020) identifies three types of mansplaining—'well, actually', straw, and speech act-confusion mansplaining. Her focus (and ours) is on the latter, which she describes as follows:

> A mansplainer [...] hears his interlocutor making a conversational move—like an assertion or hypothesis—and takes it to be a different move, one that invites him to display his expertise. In paradigm cases like this, the mansplainer takes the utterance to be a question or a request for information. This is despite the fact that the woman who is his target intends to be asserting something. The mansplainer of this type jumps in to address the woman's utterance, despite the fact that the woman, who is an expert in the relevant subject matter, took herself to be telling rather than asking (or requesting). (Johnson 2020, 5)

A hallmark of (speech act-confusion) mansplaining is a denial of uptake to a woman's utterance. Whereas the woman's performative intention is to assert a certain content (i.e., utter a factive statement), and thereby introduce said content into the common ground, the mansplainer takes her utterance to have a different illocutionary force—specifically, one that presupposes (and, if unchallenged, induces) a complete reversal of the interlocutors' conversational roles and epistemic attributes.[12] As a result of this 'confusion', that is, the man appoints himself as the resident epistemic authority, therefore appropriating the right to take the conversational floor. Correspondingly, he relegates the woman to a position of epistemic dependence, assigning her to the role of recipient (rather than originator) of informative testimony. From this position, however, the woman is kept from doing what she intended to do with her speech: her assertion is not felicitous.

Johnson notes that speech act-confusion, in and of itself, is a relatively common phenomenon (misunderstandings do happen, after all) and need not be culpable. Equally obviously, offering an explanation is certainly not harmful *per se*, and is often perfectly appropriate. Speech act-confusion mansplaining, however, is borne out of gender-based prejudice: the mansplainer mansplains because 'he reacts to the conventional procedures for asserting vs questioning in a way that is informed by the gender of the speaker' (Johnson 2020, 17). Therefore, his discursive move cannot be glossed over as innocent or non-culpable (not so easily, anyway; see (Dotson 2011) on situated ignorance). Rather, Johnson concludes, mansplaining is a form of illocutionary silencing (or disablement).[13]

[12] For an analysis of mansplaining as a form of epistemic injustice, see (Dular 2021). For an empirical study of mansplaining within a conversational analysis framework, see (Joyce *et al.* 2021).

[13] On Johnson's pluralist account, however, a mismatch between speaker intention and hearer uptake stands in the way of an illocutionary act being fully successful, but does not preclude the act from *occurring* (cf. Langton 1993).

III.2 Trolling

The literature distinguishes several, often overlapping varieties of trolling; examples include RIP and shock trolling, malicious or abusive trolling, playful or jocular trolling, concern trolling, and subcultural trolling (see e.g. Phillips 2015; DiFranco 2020; Paakki *et al.* 2021; Connolly 2022; Morgan 2022). Coursing through this diverse typology are certain recurring behavioural and structural traits, which we briefly review below.

Trolling is a complex speech act, characteristic of internet communication. Paradigmatically, the trolling agent (troll, for short) crafts a message whose content pushes socio-cultural boundaries to extremes, which he introduces into the public sphere.[14] His performative (perlocutionary) intention in so doing is twofold: provoke one type of reaction in the target, that is, the 'audience that trolling is performed *to*'; and provoke an altogether different reaction in the onlookers, that is, the audience that 'an act of trolling [...] is performed *for*' (Connolly 2022, 405).[15]

Trolling is thus designed to be divisive: its performance requires that two separate audiences understand the same locutionary content in different ways, and its success depends on the attainment of specific different reactions by each of these audiences. Thus, the target is expected to understand the troll's utterance (and the troll himself) as serious (Connolly 2022), or as expressing a sincere belief (Morgan 2022), and therefore as deserving—or indeed demanding—good faith engagement. In contrast, the onlooking audience is expected to understand that same utterance (and the troll himself) as neither serious nor sincere. Where the onlooking audience is concerned, moreover, the intended perlocutionary effect of trolling is to entertain, amuse, or—as the kids call it—acquire 'lulz'. Crucially, these expected reactions are not independent: the primary intended effect on the onlooker—entertainment—hinges on whether the target reacts as desired—for example, issuing a serious response, expressing outrage, etc.

The success of a trolling act thus fundamentally depends on getting the target to believe that *someone believes* a certain problematic, even intolerable claim, and *takes it seriously* enough to publish (utter) it. Since the troll's primary aim—provoking lulz in the onlooking audience—is achieved at the expense of (the subsidiary intended effect of affecting) the target, trolling is a manipulative and abusive act (Connolly 2022); as such, it is also (*pro tanto*) morally wrong (DiFranco 2020).

[14] Some empirical studies have shown that men are significantly more likely to carry out trolling acts. For this reason, we use the masculine grammatical pronoun for anaphoric reference purposes.

[15] According to Morgan (2022), this behavioural pattern is specific to *subcultural* trolling (SCT) rather than trolling *simpliciter*. Since we are not looking to compare or contrast different types of trolling behaviour, we gloss over these distinctions in what follows.

III.3 Comparison

The discursive behaviours of mansplainers and trolls bear some instructive local similarities to vicious devil's advocates, we think. Using STREET HARASSMENT (Section II.2) as our illustration of reference, we now bring some of these similarities into view.

STREET HARASSMENT, recall, is a first-person report of a conversational exchange within a radio segment thematically focused on women's experiences of street harassment. Having listened to the narrator's reflections on the pervasiveness of public manifestations of misogynistic attitudes, such as catcalling and 'tatcalling', and to her testimony about first-hand experiences of the same, the radio host opened the conversation to the listening audience. More specifically, the host invited (and so granted discursive permission to) a specific subset of the audience—tattooed women—to make a specific kind of conversational contribution—namely, offer testimony about their own experiences of street harassment. The first call fit neither of these descriptions, however. The caller was a man; his contribution consisted of a response, under the guise of playing devil's advocate, to the narrator's statements about the oppressive nature, and roots, of street harassment.

The male caller's contribution was inappropriate on a number of levels. First, it was inappropriate as a contribution to the conversation that was intended to take place. Following the above-described contributions by the narrator, the conversation was intended to include the testimony of tattooed women. The intended contributions were acts of telling, and the intended participants were those with the requisite epistemic standing to perform those acts. By calling into question the narrator's assertions, the male caller disrupted the conversational exchange that was intended to take place. *Disruption of conversations* is a characteristic feature of vicious devil's advocacy, mansplaining and trolling. For instance, DiFranco (2020, 939) observes that trolls characteristically violate norms constitutive of good faith public engagement, among which prescriptions against deceiving one's interlocutors as to one's conversational goals, and manipulating them for self-serving purposes, are paramount.[16]

Second, the male caller's contribution was inappropriate as a rejoinder to the narrator's testimony. On this count, STREET HARASSMENT mirrors the paradigmatic mansplaining scenario. The narrator took herself to be telling, from a position of authority on the object of that telling; the appropriate response, on the part of someone who does not occupy a similar epistemic position, is to receive the speaker's testimony (listen; perhaps update one's doxastic commitments). The male caller refused the narrator's contribution its appropriate uptake: by treating it as tantamount to an invitation

[16]See (Goldberg 2020, 43) on speakers' defeasible entitlement to claim their interlocutors' attention. Cuneo (2014) defends a normative account of speech according to which some conversational permissibility facts derive from moral facts.

to engage in a debate, or as a request for aid, his contribution—like the mansplainer's—qualified as a form of *illocutionary disablement*.[17]

Third, the male caller's contribution counts as a form of *locutionary silencing*: for as long as he stayed on the phone, those who were entitled to participate in the conversation were prevented from exercising their discursive right.

Fourth, a foreseeable perlocutionary effect of vicious devil's advocacy, and of mansplaining, is the *testimonial smothering* of other conversational participants, that is, the coerced 'truncating of one's own testimony in order to insure that the testimony contains only content for which one's audience demonstrates testimonial competence' (Dotson 2011, 244).

Fifth, all three figures *coerce their respective targets* into discursive-epistemic positions of vulnerability and dependence. Trolls and devil's advocates, in addition, place pressure on their targets to make a subsequent move (react; respond), and they do so despite lacking the requisite authority for imposing such constraints on their interlocutors.[18] In this respect, such moves may further be seen to generate *epistemically exploitative* dynamics, whereby hearers are 'required to do the unpaid and often unacknowledged work of providing information, resources, and evidence of oppression to privileged persons who demand it' (Berenstain 2016, 570). More generally, on this count too, all three figures defy some of the most basic standards governing conversational activities, holding interlocutors to reciprocal expectations of respect for one another's freedom and autonomy (Cuneo 2014; Goldberg 2020; Sbisà 2006, 2023). We will resume our parallel examination of vicious devil's advocacy, mansplaining, and trolling in Section V. First, we take stock of the implications of the foregoing for ideal analyses of devil's advocacy.

IV. 'Don't you like your ideas to be challenged?'

We have seen that diabolical devil's advocates may be imputable for locutionary, illocutionary, and perlocutionary wrongdoing. These intuitive observations naturally raise the question: What kind of speech act does a speaker perform by dint of playing devil's advocate? Surprisingly, this question has been largely neglected in the literature; we draw here on the only exception we are aware of, due to Jacobs (1989).

Jacobs identifies devil's advocacy as one among several types of argument that do not fit the 'standard' formats of pro- and contra-argumentation.[19]

[17] See also (Hazlett 2017; Wanderer 2012).

[18] The question of whether and in what sense anyone qualifies as authority in these contexts is left for future work. See (Caponetto 2022, 2023) for discussion of the authority condition in permission requests.

[19] Jacobs is operating with a conception of arguments as *complex speech acts*, originally due to van Eemeren and Grootendorst (1984).

0. *Hearer* puts forward an expressed opinion O.

Recognition conditions:

1. *Speaker* puts forward a series of assertions, $S_1, ..., S_n$, in which propositions are expressed.

[Propositional content condition]

2. Advancing $S_1, ..., S_n$ counts as an attempt by the speaker to **convince** the hearer of the **unacceptability** of O.

[Essential condition]

Correctness conditions:

3. The speaker believes that:
   (a) The hearer accepts O,
   (b) The hearer does (or will) accept $S_1, ..., S_n$,
   (c) The hearer will accept $S_1, ..., S_n$ as a refutation of O.

[Preparatory conditions]

4. The speaker believes that:
   (a) O is unacceptable,
   (b) $S_1, ..., S_n$ are acceptable,
   (c) $S_1, ..., S_n$ **refute** O.

[Sincerity conditions]

Figure 1. Felicity conditions for contra-argumentation.

In particular, Jacobs observes that despite bearing a passing resemblance to contra-argumentation, devil's advocacy does not carry the latter's characteristic persuasive force (see Fig. 1). Rather, the distinctive illocutionary point of (ideal) devil's advocacy is one of *idea-testing*: its function 'is not so much to try to convince one's interlocutor of the unacceptability of [a previously introduced assertion] *O*, as to *test for the acceptability or unacceptability of O by seeing whether one's own arguments are acceptable or unacceptable to the listener*' (Jacobs 1989, 353, emphasis added).

In fact, the speech acts of contra-argumentation and devil's advocacy differ in a number of respects, as may be seen by comparing the respective sets of felicity conditions displayed in Figs 1–2 (the former reproduces Jacobs's formulation; the latter is our own reconstruction from Jacobs's informal discussion). Both movements start out in the same way: the hearer's utterance of an assertion *O*—condition 0—is followed by the speaker's assertion of reasons

0. *Hearer* puts forward an expressed opinion O.

Recognition conditions:

1. *Speaker* puts forward a series of assertions, $S_1, ..., S_n$, in which propositions are expressed.

[Propositional content condition]

2. Advancing $S_1, ..., S_n$ counts as an attempt by the speaker to **test for the acceptability or unacceptability of** O.

[Essential condition]

Correctness conditions:

3. The speaker believes that:
   (a) The hearer accepts O.
4. The speaker *is not committed to believing* that:
   (a) The hearer does (or will) accept $S_1, ..., S_n$,
   (b) The hearer will accept $S_1, ..., S_n$ as a refutation of O.

[Preparatory conditions]

5. The speaker *is not committed to believing* that:
   (a) O is unacceptable,
   (b) $S_1, ..., S_n$ are acceptable,
   (c) $S_1, ..., S_n$ **refute** O.
6. The speaker *is committed to believing* that someone might think that:
   (a) O is unacceptable,
   (b) $S_1, ..., S_n$ are acceptable,
   (c) $S_1, ..., S_n$ **refute** O.

[Sincerity conditions]

Figure 2. Felicity conditions for devil's advocacy.

S_1,..., S_n—condition 1. The two sets of felicity conditions diverge immediately thereafter; this is as expected since, as noted, the paradigmatic point of devil's advocacy is exploratory rather than persuasive. Specifically, (ideal) devil's advocates seek to engage their interlocutor(s) in a *cooperative joint activity* of testing the robustness of O, rather than adversarially confronting them with reasons to reject O (condition 2; Jacobs 1989, 353). Accordingly, the respective correctness conditions differ on all counts but one (item 3(a) in both lists):

> In [devil's advocacy], the speaker is not committed to believing that *O* is unacceptable. Nor is the speaker committed to believing that [the premises] S_1,..., S_n are acceptable or that the hearer believes this. Nor is the speaker committed to believing that S_1,..., S_n, refute *O* or that the hearer will believe this. In fact, part of the point of devil's advocacy is to avoid the characteristic commitments of contra-argumentation. The speaker is only committed to believing that *someone might think* [and so might sincerely assert] these things, which is a pretty light commitment. (Jacobs 1989, 353-4)

The Jacobsian felicity conditions for devil's advocacy are easily seen to match the characteristic profile traits reviewed in Section II.1. The ideal devil's advocate, recall, deliberately 'takes a position she does not necessarily agree with' (Johnson 2018a, 99) and proceeds to defend it (condition 1). Knowing that the position is not shared by her interlocutor (condition 3), the devil's advocate thereby occupies the role of critical opponent in the unfolding exchange (Stevens and Cohen 2019; condition 1). Her doing so counts as an attempt to 'explore [the hearer's position, *O*] further' (Johnson 2018a, 99; condition 2), by countenancing potential challenges to its justificatory basis (conditions 4–6).

With the felicity conditions of ideal devil's advocacy in hand, let us return once more to the conversational exchange between Male Caller and Radio Guest (as we'll label them for ease of reference) in Street harassment. This will help bring into view the explanatory limitations of ideal accounts, and provide an additional diagnostic clue as to how what goes on—and goes wrong—in this and similar scenarios.

IV.1 Limitations of ideal analyses

Catcalls are oppressive speech acts. This is so whether or not their semantic content is derogatory; even when they superficially present as benign (e.g. 'Nice curves!'), catcalls locutionarily, illocutionarily, and perlocutionarily demean their targets (McDonald 2022; Hesni 2018). And, since catcalls reproduce and make salient structures of gender-based oppression (Simpson 2013), and are overwhelmingly addressed by men towards members of non-dominant genders (in particular, women), they are recognizable expressions of sexist attitudes (see also Goldberg 2020, 47–8). We take this much as a given in what follows.

Recall now that Male Caller purported to test ('for the sake of argument') radio guest's claim that 'tatcalls', like catcalls, are a form of public harassment rooted in patriarchal oppression. For ease of analysis, let's take the starting point of the exchange to be Radio Guest's utterance of the following:

> (O) Catcalls are a symptom of patriarchal oppression.

Let's also assume that Male Caller purported to test O (condition 2) by offering reasons (condition 1) which '*someone might think*' count as reasons against O and in favour of C, instead (condition 6):

(C) Catcalls are innocent compliments (and not a symptom of patriarchal oppression).

Earlier, we argued that Male Caller's contribution was problematic insofar as it constituted a form of silencing—of locutionary, illocutionary, and perlocutionary varieties (Section III.3). Notice now that this assessment did not, in any way, depend on the *content* of Male Caller's utterance (i.e., C). Had Male Caller given a rendition of Bohemian Rhapsody, he still would have disrupted the conversation that was intended and expected to take place; he would still have counted as having denied appropriate uptake to Radio Guest's testimony; and he would still have prevented listeners from phoning in to offer their testimony. In this (counterfactual) case, though, even ideal analyses would have found fault with Male Caller's utterance because *some* constraints on locutionary content must be met in order for any speech act to get off the ground. So, uttering 'I promise I was at the bar last night' does not accomplish the speech act of promising: to count as a promise, the speaker's locution must refer to an act she intends to perform in the future. Similarly, uttering 'I'm just playing devil's advocate here, but—is this the real life? Is this just fantasy? Caught in a landslide, no escape from reality...' does not accomplish the act of playing devil's advocate since the lyrics of Bohemian Rhapsody are not assertions.

Male Caller's actual utterance clearly passed this minimal locutionary threshold. Did it also fulfil the remaining felicity conditions of devil's advocacy? At first glance, it would seem that this question could only be settled by accessing Male Caller's cognitive state—since, as noted, the felicity conditions of devil's advocacy and contra-argumentation are indiscernible with respect to propositional content. This is far from anomalous: plenty of locutions may be used to perform very different illocutionary acts depending on what speakers believe, intend, etc.[20] Thus, the locution 'I'll make you sashimi for dinner' may alternately count as a promise or a threat depending on whether the speaker believes the hearer enjoys sashimi or is repulsed by it. In the absence of evidence as to what a speaker believes, by contrast, the illocutionary force of such an utterance may be underdetermined.

In the case of devil's advocacy, however, the threat of illocutionary underdetermination is allayed by *the preamble*, which is specifically and exactly designed to signal that the speaker is contributing to the exchange under the premise of (open) insincerity. In effect, the preamble simultaneously brings to salience one set of felicity conditions (idea-testing), and relegates another to the back seat (refutation).

Had Male Caller's contribution featured the bare (non-embedded) assertive C, in response to Radio Guest's utterance of O, he would count as having engaged in an adversarial act of contra-argumentation. However, *the preamble*

[20]It has also been argued that speakers may perform multiple illocutionary acts via a single locutionary utterance. For discussions of illocutionary pluralism and illocutionary relativism, respectively, see (Lewiński 2021; Johnson 2023).

changes everything—so extant analyses imply. Two upshots are especially important; we'll see just how important they are in the next section. First, the sincerity-suspending effects of the preamble make it so that the utterance of C does not count as an assertion of C; nor, *a fortiori*, as reflecting Male Caller's actual (private) commitments. Second, Male Caller counts as performing a speech act whose illocutionary point is idea-testing, whose overall spirit is cooperative, and which is intended to benefit Radio Guest (and perhaps the listening audience). In effect, Male Caller's competent performance of the act of devil's advocacy automatically colours his contribution with a presumptive tint of virtuosity: all he wanted to do was kick some ideas around, for the sake of argument.

This is unsatisfactory. Uttering C under a premise of open insincerity is not inherently wrong; indeed, it is politically and epistemically imperative that there be room to sometimes discuss morally problematic, even distasteful views. However, nor is it inherently permissible, much less praiseworthy, to do so unqualifiedly—at any point, in any conversation, defying any other discursive-epistemic norms governing the exchange (e.g. pertaining to uptake, turn-taking, and extending testimonial evidence). Ideal analyses are unequipped to determine where the boundaries of appropriate devil's advocacy lie, because they take the preamble to effectively dispense with the need to consult the larger conversational context. However, treating the preamble as a miraculous panacea is misguided, in much the same way that it is misguided to insist that the right to free speech needs no qualifications.[21] And it is unsatisfactory, because by glossing over the dramatic illocutionary effects of the preamble, no meaningful distinction can be made between genuinely virtuous devil's advocates and their vicious counterparts. The next section drives home the importance of tracing these boundaries by looking more closely at the powerful effects of the preamble on the dynamics of conversation.

V. The weaponization of illocutionary force

V.1 Figleaves

Above, we said that C, as a pure assertive, is a recognizable expression of a sexist attitude. In fact, C is recognizable more specifically as an expression of what is sometimes denominated 'modern' sexism (Regnier-Bachand 2015): in contrast with expressions of overt hostility towards women (more generally: towards non-dominant genders), C denotes a predisposition to assent to, or justify, at least some forms of gender-based discrimination.

[21] We thank an anonymous reviewer for suggesting that we flag the connections between our discussion and the free speech debate. Though length limitations prevent us from elaborating further, we hope to return to this matter in future work.

Sexist attitudes may be mapped by measures of hostility towards women, and according to whether they present as overtly discriminatory or as covert, 'benevolent' forms of sexism. Sexist attitudes may also be mapped according to the perceived tolerability of their public expression. In recent work, Saul (2021) argues that given the widespread endorsement of a fairly generic norm of gender equality, it is reasonable to expect that most people will seek to avoid being associated with recognizable expressions of sexism. Similar considerations may explain the fact that bare racist assertives are also (relatively) rarely found in public discursive contexts, since they are difficult to reconcile with a matching norm of racial equality (Saul 2017, 2021).

The endurance of racial and gender equality norms makes it so that overtly racist and sexist utterances tend to be poorly tolerated in the public realm, even while attitudes of resentment towards members of racial minorities (respectively: non-dominant genders) remain widespread.[22] It also makes it more likely that individuals will tend to avoid introducing into the record content that could easily warrant attributions of, or association with, racism or sexism. In contrast, tolerance seems to be higher when expressions of racial and gender resentment are accompanied by additional overt markers that allow speakers (and hearers) to disavow attributions of racism or sexism. Saul terms these *racial* and *gender figleaves*, respectively: utterances that 'provide a small bit of cover for something that is unacceptable to display in public' (Saul 2017, 98).

Figleaves are identified functionally, as utterances 'which (for some portion of the audience) [block] the conclusion that (a) some other utterance, R, is racist [or sexist]; or (b) the person who uttered R is racist [or sexist]' (Saul 2021, 161). Alongside a variety of such utterances operating at a direct propositional level (e.g. 'I'm not a racist, but...', 'I have great respect for women, but...'), Saul identifies a further type of figleaf—the *force figleaf*—that we think helps diagnose at least some vicious uses of devil's advocacy. As before, the hallmark functional trait of a force figleaf is that of blocking the inference that either the speaker, or the utterance, are racist (sexist). But here, the screening-off effect is achieved by *changing the way that the audience understands the speech act being performed* by the speaker. That is, force figleaves—when successful—modify the on-record illocutionary force of the relevant speech act.

One of the examples discussed by Saul is drawn from the vast repertoire of Donald Trump's controversial public statements: 'When the revelation of [the Access Hollywood tapes] led to an uproar, just before the election, Trump responded by saying that the comments were merely "locker room talk"' (Saul 2021, 170). Trump's locker-room-shaped hedge qualifies as a force figleaf: its

[22] See also (Mendelberg 2001; Khoo 2017). Expressions of racial and gender resentment appeared in Affirmative action ('... whether maybe I really hadn't deserved admission to my law school') and Wage gap ('Women take more time off of work'), respectively.

(intended) function was to distance Trump from the content of his prior, overtly sexist utterances. To achieve this purpose, Trump introduced a presupposition to the effect that uttering overtly sexist remarks, while unacceptable in ordinary circumstances, *is however permissible when embedded in a locker-room conversational context*. For, in such contexts, specific felicity conditions are in place that rule out the face-value interpretation of those utterances. Thus, the intended effect of Trump's remark was to retroactively modify the intended illocutionary force of his previous speech act, thereby allowing a portion of his audience to avert the otherwise inevitable conclusion that Trump's assertion, and Trump himself, were sexist.

V.2 Two routes to plausible deniability

The function of force figleaves, and figleaves more generally, is to offer speakers (and hearers) an out: if the figleaf is successful, speakers (and hearers) will be in a position to plausibly deny undesirable, and otherwise inevitable, attributions of racism or sexism. In practice, force figleaves are intended to achieve this goal by 'relocating' the problematic utterance in a sincerity-suspending context: one whose felicity conditions drive a wedge between the embedded locutionary content and the doxastic commitments that may reasonably be attributed to the speaker on the basis of the utterance in question. Familiar hedges such as 'I was only joking', 'I'm just curious/asking a question', or 'I'm just quoting X on this', offer a similar promise of plausible deniability; in all these cases, the fact that the respective speech acts lack the sincerity conditions characteristic of assertions stands in the way of ascribing utterer beliefs on the basis of uttered content, and of treating the latter as a putative truth claim.[23]

Notice, now, that trolling and mansplaining may also count on strategies for securing plausible deniability via the manipulation of illocutionary force. The specifics are worth spelling out in some detail, both because they are interesting in their own right and because they will help illuminate certain distinctive traits of devil's advocacy as an especially powerful variant of this strategy.

Force figleaves, we've seen, trade on a suspension of speaker sincerity. So too do trolling acts, as we saw in Section III.2. Figleaves, however, are overtly marked: they are designed to be in full public view. On the contrary, the whole point of trolling acts is to ensure that an utterance is simultaneously understood as sincere by one part of the audience (target) and as insincere by the

[23]The promise of plausible deniability offered by such hedges is not always an enthusing one, note. Boogaart *et al.* (2021 2022) discuss the prospects of various such defensive strategies as tickets to plausible deniability. See also (Camp 2022).

rest (onlookers).[24] A precondition of successful trolling, then, is the absence of any explicit markers that might reveal the troll's performative intention to the target, *while simultaneously ensuring* that the onlookers understand that his locutionary act is *not* subject to a condition of sincerity. This means that a pathway to plausible deniability *relative to locutionary content* is an inbuilt feature of the complex speech act of trolling. In this sense, we suggest that trolling is functionally comparable to the deployment of a force figleaf—again, relative to the locutionary content of the troll's utterance.

The qualification 'relative to locutionary content' is crucial: what does not seem to be available to the troll is a path to plausible deniability relative to *illocution*; nor, *a fortiori*, relative to his overall moral standing. For, the only way for a trolling agent to avoid condemnation would be to dismiss (qualify) his own actions as 'mere trolling'. However, it is unlikely that a troll would do this, since he would thereby cancel an essential felicity condition of his own speech act (see also Morgan 2022, 16). And it is far from clear that a troll could do this even if he wanted to, since there is no acknowledged morally neutral counterpart to the act of trolling.[25] Thus, while the presupposition of insincerity introduces a wedge (in principle, at least) between uttered content and utterer's private commitments, there is no corresponding consideration that could plausibly shield the troll from reproach *qua* agent: in particular, there is no plausible redeeming construal of the troll's deliberately coercive intention towards his target.

In this latter respect, trolling acts are unlike figleaf defences, which typically block (or purport to block) attributions of moral badness to the utterer if they block attributions of moral badness to the utterance, and vice versa.[26] Trolling is also unlike mansplaining, in this and almost every respect; this is unsurprising given that the presuppositions constitutive of mansplaining are all relative to the mansplainer's epistemic and moral standing, rather than to the content of his contribution. Mansplaining is problematic not because of what the speaker says, but because—exactly and only because—of what the speaker *does*. And what the speaker does is manipulate illocutionary force: he replaces the felicity conditions of his interlocutor's actual speech act (e.g. giving testimony) with a different set of conditions (e.g. the felicity conditions of asking for help, requesting advice or information).

[24]More fully: the troll's objective is to make it so that the target audience (the trollee, as it were) comes to believe that *someone* sincerely believes the utterance in question. An additional precondition of successful trolling is that the target should not conceive of this 'someone' as a troll.

[25]In contrast, the fact that joking is generally considered to be a morally neutral activity suggests that 'I'm just joking' defences will be at least marginally stronger in this respect.

[26]An interesting question is whether anything changes when an utterance is explicitly marked as an act of trolling by a third party—as when the *LA Times* dismissed one of Trump's many racist rants as 'mere trolling': https://www.latimes.com/opinion/editorials/la-ed-trump-aoc-squad-ilhan-bigoted-tweets-20190714-story.html.

Notice that this same manipulation opens the door to a powerful redeeming presupposition, to the effect that the mansplainer is performing a (morally and epistemically) positively valued speech act, the key upshot of which will be to improve his interlocutor's epistemic standing. The mansplainer thus makes the 'bold' move of trying to pass off a harmful act as one that is epistemically laudable (explanation is both an epistemic end in itself, and a vehicle to other epistemic ends); and potentially, also morally so (offering aid, including epistemic aid, reflects positively on one's character, and may even be a duty).

Notice also that the above presupposition invites the conclusion that the act being performed by the mansplainer is non-culpable. Plausible deniability may thus be established by trading the stronger presupposition for the weaker—once again, by massaging illocutionary force:

> The man wasn't doing anything pernicious, he might argue, he was just ϕ-ing, where ϕ-ing is something we all occasionally non-culpably do. (Johnson 2020, 6).[27]

Thus, the fact that explanations, clarifications and answers may be praiseworthy discursive contributions, combined with the fact that speech act-confusion may be non-culpable, jointly open a path to possible and plausible deniability.[28]

V.3 Diabolical devil's advocates

Some speech acts—such as trolling—harm as a result of the interaction between illocutionary force and locutionary content. Others—such as mansplaining—harm in virtue of illocutionary force alone, regardless of content. Ultimately, moral assessments always reflect back on speech *agents*; but the distinction helps track the different routes agents may take to establish non-culpability. In turn, the promise of any defensive option may vary greatly across socio-cultural contexts and over time (Saul 2021), but also according to standing pragmatic conventions and particularized contextual details (Boogaart *et al.* 2021). Even so, it seems likely that some defensive strategies are *ceteris paribus* stronger than others (Boogaart *et al.* 2022). Indeed, one of the morals emerging from our analysis of trolling and mansplaining is that the existence of a nearby, 'lookalike' speech act that is recognized as non-culpable

[27] Tucker Carlson seemingly employed this line of defence when he responded, to the charge of mansplaining issued by a guest on his show: 'I'm not mansplaining, I'm saying something that's obviously true' (Joyce *et al.* 2021, 509).

[28] Requests threaten both the listener's negative face—freedom of action and from imposition—and the speaker's positive face—desire for approval of her self-image (Brown and Levinson 1987). The mansplainer can reclaim virtuosity *precisely* because requests for (epistemic) help (e.g. alternative explanation, information) are acts that restrict the requestee's personal freedom *for the benefit of the requester*. As Dular (2021, 15) puts it: 'Although the explaining is meant to look innocuous, such "helping" is really hurting, a self-interested move thinly veiled as self-sacrifice.'

may significantly improve the prospects of a speaker's attempt to disavow socially costly commitments.

Where does devil's advocacy stand, in all of this? In a uniquely favourable position, we now argue. Consider the following utterance, loosely adapted from Affirmative action:

> (A^*) I'm just playing devil's advocate here, but let's consider whether affirmative action policies are unfair to whites.

A^* is uttered in response to the narrator's (Law Graduate) act of offering testimony; it is uttered by the narrator's friend (White Friend), who was the recipient of that testimony. A^* embeds an assertion, A, that is recognisable as an expression of racial resentment; a bare utterance of A would betray a very weak commitment to the norm of racial equality on the utterer's part (see Section V.1). In uttering A^*, in contrast, White Friend signals that he is voicing A *on behalf of someone* (actual or imagined) *other than himself*. This deflection of commitments is effected via the preamble: White Friend is on record as having merely put A up for joint discussion. In the same fell swoop, the preamble also affects Law Graduate's deontic profile (see e.g. Sbisà 2006). Minimally: it obliges her to attend to his contribution (Goldberg 2020); it compels her to respond to White Friend's attempt to engage in a joint idea-testing activity; it precludes her from (justifiably) ascribing a racially resentful attitude to White Friend; and it precludes her from (justifiably) evaluating White Friend's moral profile on the grounds of his uttering A.

In addition, given that boilerplate expectations of cooperativeness are ostensibly warranted in this particular context, the preamble *also* entitles Law Graduate to believe that White Friend regards A as *worthy of consideration*; as relevant to the unfolding exchange; and as requiring discussion there and then.

Being worthy of consideration is not an especially high standard. However, it does rule out a sizeable quantity of things: there is a sizeable quantity of beliefs that are generally regarded as being beyond the pale. If Mendelberg, Saul, and others are right about the widespread acceptance of the racial equality norm, then racist beliefs are among the beliefs that are generally regarded as beyond the pale; correspondingly, introducing such beliefs into the conversational record under a premise of speaker sincerity (e.g. via assertion) is generally regarded as unacceptable. From the fact that White Friend embeds A in the preamble, we may glean that he is sensitive to the racial equality norm. However, we may also glean that he is committed to a worryingly weak reading of the norm: by embedding A in the preamble, White Friend simultaneously signals that while he rejects A, he *also* regards it as worthy of consideration.

Thus, the preamble blocks, or purports to block, the inference to the conclusion that the speaker is racist (sexist); and it blocks, or purports to block, the inference to the conclusion that the embedded utterance is racist (sexist). Since

it blocks these inferences by intervening on illocutionary force, the preamble is functionally recognizable as a force figleaf. Two things follow from this. First, *qua* figleaf, the preamble affords the speaker a putative plausible deniability defence relative to the introduction of problematic content into the conversational record. Second, as with figleaves more generally, abuses of devil's advocacy are liable to contribute to the gradual shifting of boundaries of what it is considered permissible to say in public—what one can say without fear of incurring social costs, or challenges to one's moral or epistemic profile.[29] And, we can add, of *how* one can do this: successful illocutionary manipulation begets a normalization of abusive agency in the conversational sphere.[30]

On each of these counts, *successful* deniability is key; and the availability of multiple defensive strategies enhances the prospects of success, we've argued. Diabolical devil's advocates are well placed in both respects. They may appeal to a defence relative to *locution*, as we've just seen. And they may appeal to a defence relative to *illocution*: like mansplainers, they may insist that what they were doing was non-problematic. The devil's advocate's defence, however, is stronger than the mansplainer's. The latter boils down to 'Oops! I misunderstood what was happening here. (Still, it doesn't hurt to repeat what we both know.)' The mansplainer's defence is premised on his acknowledgement that he has been caught out doing something wrong; this drives him to reach for the felicity conditions of a nearby, positively valued illocutionary act to rehabilitate himself.

In contrast, diabolical devil's advocates quite literally *wear their plausible deniability on their sleeve*: they do not need to reach for a morally preferable replacement for what they were doing, because what they were doing needs no such replacement. Any charge of wrongdoing merely reveals that the audience misunderstood what was happening; instead of acknowledging their own speech act-confusion, however, they resort to maligning the speaker. There was no silencing, no coercion, no refusal of testimony, no manipulation: the speaker was just playing devil's advocate.

On the whole, devil's advocacy is uniquely placed, among known discursive moves, to offer cover for, and normalize, abusive speech agency. Like other force figleaves, it exploits its in-built suspension of sincerity to cover for the introduction of problematic locutionary content. Like mansplaining, it may appeal to speech act-confusion to cover for problematic illocutionary behaviour, and for displaying a manifest disregard for an interlocutor's epistemic profile. Like trolling, it gestures towards a purportedly worthy trade-off

[29] See (Saul 2017; McGowan 2019). See also (Van Dijk 1992).

[30] Similar detrimental effects on the normativity of public discourse are imputable to trolling and mansplaining. For instance, Johnson and Dular both argue that mansplaining recreates unjust socio-political structures within the epistemic domain. The difficulty of challenging such conduct foregrounds its normalization, thus reinforcing existing social inequalities.

(epistemically beneficial idea-testing) to cover for the problematic perlocutionary effects of the act—*viz*., placing unjust pressure on the respective target to either respond, or else allow the problematic content to remain unchallenged (so targets are left in a double bind; see e.g. Hirji 2021).

Thus, devil's advocacy shares several traits of other abusive discursive practices. However, while the latter may count, at best, on defences that claim permissibility and lessened (or non-) culpability, the former counts on all of these things *and claims virtuousness*, to boot.

VI. Concluding Remarks

Recently, social epistemologists have sought to substantiate the intuition that we have an interpersonal epistemic duty 'to object to things that people say', as a way of providing epistemic aid and promoting epistemic goods in others (Johnson 2018b; Lackey 2020; Terzian and Corbalán 2021). In general, this duty is *imperfect*. Thus, it may be heightened when, for instance, an assertion introduced into the record expresses evidence-resistant beliefs (such as racist and sexist beliefs).[31] Moreover, as with imperfect duties more generally, we can do more than what our share demands; and such supererogatory excess may have moral value (Schroeder 2014).

Now, while there may not be a corresponding epistemic obligation to speak up on behalf of others (Johnson 2018a), it seems clear that there can be *epistemic value* in doing so. The theoretical characterisations of ideal devil's advocacy presented in Section II.1 bear witness to the naturalness of this intuition: by taking on the role of devil's advocate we can help others improve their epistemic standing, and promote the pursuit of epistemic goods more generally. A tempting continuation of this thought is that in this capacity, too, doing more will make us better: we may reach for new heights of virtuosity by giving voice to others' beliefs, even—or especially—ones we don't ourselves subscribe to.

However, this conclusion rests on a fundamental equivocation: just as giving money to the KKK does not count as an act of charity (Rainbolt 2000), so too giving voice to racist and sexist beliefs does not count as an act of epistemic aid. In a way, the observed patterns of problematic abuses of devil's advocacy in real-world conversational contexts may be seen as resting on, and even exploiting, this equivocation. Given the strong default presumption of virtuosity attached to the profile, however, it is an equivocation that is especially difficult to dismiss. As a result, and instead, we may witness the emergence of devil's advocates who give voice to oppressive views, thus skewing the distribution of higher-order evidence in dangerous directions (e.g. by 'treating racism as

[31] See (Puddifoot 2021; Cella *et al.* 2022).

a political disagreement around which two opposing viewpoints can reasonably form') and forcing recipients into positions of epistemic and discursive vulnerability—all in the name of 'open inquiry'.

A natural continuation of the present investigation will require addressing a pressing prescriptive question: What should be done? For instance, should the targets of vicious devil's advocacy *respond*, or should they *disengage*? A few cursory observations suffice to show that these are complex issues, deserving of considered separate treatment. For instance, responding may backfire: engagement is naturally interpreted as conveying assent to the presupposition that the uttered content is a proper topic of good faith, reasonable disagreement, and this in turn may boost the credibility of the proffered beliefs. Moreover, it may unfairly burden the targets themselves, who typically occupy structurally disadvantaged socio-epistemic positions in which they are already unfairly burdened. Yet, refusing to engage with the resident devil's advocates may also end up damaging their targets, since doing so may invite attributions of closemindedness and uncooperativeness.

The challenge of coming up with strategies to resist vicious discursive behaviour, without thereby reinforcing the oppressive structures to which targets are already subjected, remains a pressing one.[32] By shining a light on the dark side of devil's advocacy, we hope to have contributed to this project.

Acknowledgments

We thank audiences at ArgLab-IFILNOVA, ECA2022-European Conference on Argumentation, CIbA2023-Congreso Iberoamericano de Argumentación, Álvaro Domínguez-Armas, Amalia Haro Marchal, and two anonymous reviewers for helpful feedback on earlier versions of this paper at various stages of its development. Terzian acknowledges the financial support of the Portuguese Foundation for Science and Technology (FCT) through grant UIDP/00183/2020.

References

Aikin, S. F. and Clanton, J. C. (2010) 'Developing Group-Deliberative Virtues', *Journal of Applied Philosophy*, 27/4: 409–24. https://doi.org/10/dwvt68

Beatty, J. and Moore, A. (2010) 'Should We Aim for Consensus?', *Episteme*, 7/3: 198–214. https://doi.org/10.3366/E1742360010000948

Berenstain, N. (2016) 'Epistemic Exploitation', *Ergo*, 3/22: 569–90. https://doi.org/10.3998/ergo.12405314.0003.022

[32]The growing literature on counterspeech holds great promise on this count; see for example, Langton (2018), Caponetto and Cepollaro (2022).

Boogaart, R., Jansen, H., and van Leeuwen, M. (2021) '"Those are your Words, not Mine" Defence Strategies for Denying Speaker Commitment', *Argumentation*, 35/2: 209–35.
Boogaart, R., Jansen, H., and van Leeuwen, M. (2022) '"I was only Quoting": Shifting Viewpoint and Speaker Commitment', In: L. R. Horn (ed.) *From Lying to Perjury: Linguistic and Legal Perspectives on Lies and Other Falsehoods*, vol. 3, pp. 113–38. Walter de Gruyter.
Brohinsky, J., Sonnert, G., and Sadler, P. (2022) 'The Devil's Advocate Dynamics of Dissent in Science Education', *Science & Education*, 31: 575–96. https://doi.org/10.1007/s11191-021-00264-5
Brown, P. and Levinson, S. C. (1987) *Politeness: Some Universals in Language Usage*. Cambridge: CUP.
Camp, E. (2018) 'Insinuation, Common Ground and the Conversational Record', In: D. Fogal, D. W. Harris and M. Moss (eds.) *New Work on Speech Acts*, vol. 40, pp. 40–66. OUP.
Camp, E. (2022) 'Just Saying, just Kidding: Liability for Accountability-Avoiding Speech in Ordinary Conversation, Politics and Law', In: L. R. Horn (ed.) *From Lying to Perjury: Linguistic and Legal Perspectives on Lies and Other Falsehoods*, pp. 227–258. Walter de Gruyter.
Caponetto, L. (2022) 'Accommodated Authority: Broadening the Picture', *Analysis*, 82/4: 682–92. https://doi.org/10.1093/analys/anac049
Caponetto, L. (2023) 'The Pragmatic Structure of Refusal', *Synthese*, 201/6: 1–19. https://doi.org/10.1007/s11229-023-04177-4
Caponetto, L. and Cepollaro, B. (2022) 'Bending as Counterspeech', *Ethical Theory and Moral Practice*, 26/4: 577–93. https://doi.org/10.1007/s10677-022-10334-4
Cella, F., Marchak, K. A., Bianchi, C., and Gelman, S. A. (2022) 'Generic Language for Social and Animal Kinds: An Examination of the Asymmetry Between Acceptance and Inferences', *Cognitive Science*, 46/12: e13209. https://doi.org/10.1111/cogs.13209
Connolly, P. J. (2022) 'Trolling as Speech Act', *Journal of Social Philosophy*, 53/3: 404–20. https://doi.org/10.1111/josp.12427
Cuneo, T. (2014) *Speech and Morality: On the Metaethical Implications of Speaking*. Oxford: OUP.
DiFranco, R. (2020) 'I Wrote This Paper for the Lulz: The Ethics of Internet Trolling', *Ethical Theory and Moral Practice*, 23/5: 931–45. https://doi.org/10.1007/s10677-020-10115-x
Dotson, K. (2011) 'Tracking Epistemic Violence, Tracking Practices of Silencing', *Hypatia*, 26/2: 236–57.
Dular, N. (2021) 'Mansplaining as Epistemic Injustice', *Feminist Philosophy Quarterly*, 7/1. https://doi.org/10.5206/fpq/2021.1.8482
Duran, N. D. and Fusaroli, R. (2017) 'Conversing with a Devil's Advocate: Interpersonal Coordination in Deception and Disagreement', *PloS One*, 12/6: e0178140.
Goffman, E. (1959) *The Presentation of Self in Everyday Life*. New York, NY: Garden City.
Goldberg, S. (2020) *Conversational Pressure: Normativity in Speech Exchanges*. Oxford: OUP.
Hazlett, A. (2017) 'On the Special Insult of Refusing Testimony', *Philosophical Explorations*, 20/sup1: 37–51. https://doi.org/10.1080/13869795.2017.1287293
Hesni, S. (2018) 'Illocutionary Frustration', *Mind*, 127/508: 947–76.
Hirji, S. (2021) 'Oppressive Double Binds', *Ethics*, 131/4: 643–69. https://doi.org/10.1086/713943
Jacobs, S. (1989) 'Speech Acts and Arguments', *Argumentation*, 3/4: 345–65. https://doi.org/10.1007/BF00182603
Johnson, C. R. (2018a) 'For the Sake of Argument: The Nature and Extent of Our Obligation to Voice Disagreement', In: *Voicing Dissent*, pp. 97–108. New York, NY: Routledge. https://doi.org/10.4324/9781315181189-7
Johnson, C. R. (2018b) 'Just Say No: Obligations to Voice Disagreement', *Royal Institute of Philosophy Supplement*, 84: 117–38. https://doi.org/10.1017/S1358246118000577
Johnson, C. R. (2018c) *Voicing Dissent: The Ethics and Epistemology of Making Disagreement Public*. New York, NY: Routledge.
Johnson, C. R. (2020) 'Mansplaining and Illocutionary Force', *Feminist Philosophy Quarterly*, 6/4. https://doi.org/10.5206/fpq/2020.4.8168
Johnson, C. R. (2023) 'Illocutionary Relativism', *Synthese*, 202/3: 1–18.
Joyce, J. B., Humă, B., Ristimäki, H.-L., Ferraz de Almeida, F., and Doehring, A. (2021) 'Speaking Out Against Everyday Sexism: Gender and Epistemics in Accusations of "Mansplaining"', *Feminism & Psychology*, 31/4: 502–29.

Khoo, J. (2017) 'Code Words in Political Discourse', *Philosophical Topics*, 45/2: 33–64.
Lackey, J. (2020) 'The Duty to Object', *Philosophy and Phenomenological Research*, 101/1: 35–60.
Langton, R. (1993) 'Speech Acts and Unspeakable Acts', *Philosophy & Public Affairs*, 22/4: 293–330.
Langton, R. (2018) 'Blocking as Counter-Speech', In: D. Fogal, D. W. Harris and M. Moss (eds.) *New Work on Speech Acts*, pp. 144–164. Oxford: OUP. https://doi.org/10.1093/oso/9780198738831.003.0006
Lewiński, M. (2021) 'Illocutionary Pluralism', *Synthese*, 199/3–4: 6687–714.
Lewis, D. (1979) 'Scorekeeping in a Language Game', *Journal of Philosophical Logic*, 8: 339–59.
McDonald, L. (2022) 'Cat-Calls, Compliments and Coercion', *Pacific Philosophical Quarterly*, 103/1: 208–30. https://doi.org/10.1111/papq.12385
McGowan, M. K. (2019) *Just Words: on Speech and Hidden Harm*. Oxford: OUP.
Mendelberg, T. (2001) *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. NJ: Princeton University Press.
Mill, J. S. (1859) *On Liberty*. London: John W. Parker & son.
Morgan, A. (2022) 'When Doublespeak Goes Viral: A Speech Act Analysis of Internet Trolling', *Erkenntnis*, 88: 3397–417. https://doi.org/10.1007/s10670-021-00508-4
Nemeth, C., Connell, J., Rogers, J., and Brown, K. (2001a) 'Improving Decision Making by Means of Dissent', *Journal of Applied Social Psychology*, 31/1: 48–58. https://doi.org/10.1111/j.1559-1816.2001.tb02481.x
Nemeth, C. J., Brown, K. S., and Rogers, J. D. (2001b) 'Devil's Advocate versus Authentic Dissent: Stimulating Quantity and Quality', *European Journal of Social Psychology*, 31/6: 707–20. https://doi.org/10.1002/ejsp.58
Paakki, H., Vepsäläinen, H., and Salovaara, A. (2021) 'Disruptive Online Communication: How Asymmetric Trolling-like Response Strategies Steer Conversation off the Track', *Computer Supported Cooperative Work*, 30/3: 425–61.
Phillips, W. (2015) *This Is Why We Can't Have Nice Things: Mapping the Relationship between Online Trolling and Mainstream Culture*. Boston, MA: MIT Press.
Puddifoot, K. (2021) *How Stereotypes Deceive Us*. Oxford: OUP.
Rainbolt, G. (2000) 'Perfect and Imperfect Obligations', *Philosophical Studies*, 98/3: 233–56.
Regnier-Bachand, C. M. (2015) *Sexism and Women: The Implications of Female Gender Resentment*. Master of Arts, Department of Political Science at the University of Central Florida, Orlando.
Sacks, H. and Schegloff, E. A. (1979) 'Two Preferences in the Organization of Reference to Persons in Conversation and their Interaction', in P. George (ed) *Everyday Language: Studies in Ethnomethodology*, pp. 15–21. New York: Irvington.
Saul, J. M. (2017) 'Racial Figleaves, the Shifting Boundaries of the Permissible, and the Rise of Donald Trump', *Philosophical Topics*, 45/2: 97–116. https://doi.org/10.5840/philtopics201745215
Saul, J. M. (2021) 'Racist and Sexist Figleaves', in J. Khoo and R. K. Sterken (eds) *The Routledge Handbook of Social and Political Philosophy of Language*, pp. 161–178. New York, NY: Routledge.
Sbisà, M. (2006) 'Communicating Citizenship in Verbal Interaction: Principles of a Speech Act Oriented Discourse Analysis', In: H. Hausendorf and A. Bora (eds.) *Analysing Citizenship Talk*. Amsterdam: John Benjamins Publishing Company.
Sbisà, M. (2023) *Essays on Speech Acts and Other Topics in Pragmatics*. Oxford: Oxford Academic. https://doi.org/10.1093/oso/9780192844125.001.0001
Schroeder, S. A. (2014) 'Imperfect Duties, Group Obligations, and Beneficence', *Journal of Moral Philosophy*, 11/5: 557–84.
Searle, J. R. (1976) 'A Classification of Illocutionary Acts', *Language in Society*, 5/1: 1–23.
Simpson, R. M. (2013) 'Un-ringing the Bell: McGowan on Oppressive Speech and the Asymmetric Pliability of Conversations', *Australasian Journal of Philosophy*, 91/3: 555–75.
Stalnaker, R. (2002) 'Common Ground', *Linguistics and Philosophy*, 25/5-6: 701–21.
Stevens, K. and Cohen, D. H. (2019) 'Devil's Advocates are the Angels of Argumentation', In: *Reason to Dissent: Proceedings of the 3rd European Conference on Argumentation*, vol. 2, pp. 161–174. London: College Publications.
Terkourafi, M. (2015) 'Conventionalization: A New Agenda for Im/politeness Research', *Journal of Pragmatics*, 86: 11–18.

Terzian, G. and Corbalán, M. I. (2021) 'Our Epistemic Duties in Scenarios of Vaccine Mistrust', *International Journal of Philosophical Studies*, 29/4: 613–40. https://doi.org/10.1080/09672559.2021.1997399

Van Dijk, T. A. (1992) 'Denying Racism: Elite Discourse and Racism', *Discourse and Society*, 3/1: 87–118.

van Eemeren, F. H. and Grootendorst, R. (1984) *Speech Acts in Argumentative Discussions: A Theoretical Model for The Analysis of Discussions Directed towards Solving Conflicts of Opinion*. vol. 1. New York, NY: De Gruyter Mouton.

Wanderer, J. (2012) 'Addressing Testimonial Injustice: Being Ignored and Being Rejected', *The Philosophical Quarterly*, 62/246: 148–69. https://doi.org/10.111/j.1467-9213-2011.712.x