Philosophy & Technology (2024) 37:35
https://doi.org/10.1007/s13347-024-00725-z

RESEARCH ARTICLE

Wrongful Rational Persuasion Online

Thomas Mitchell[1] · Thomas Douglas[1,2]

Received: 3 October 2023 / Accepted: 20 February 2024 / Published online: 28 February 2024

Abstract

In this article, we argue that rational persuasion can be a *pro tanto* wrong and that online platforms possess features that are especially conducive to this wrong. We begin by setting out an account of rational persuasion. This consists of four jointly sufficient conditions for rational persuasion and is intended to capture the core, uncontroversial cases of such persuasion. We then discuss a series of wrong-making features which are present in methods of influence commonly thought of as *pro tanto* wrong, such as manipulation and paternalism. It is next shown that these wrong-making features are also present in a range of cases that are, by the jointly sufficient conditions already established, rational persuasion, and so some forms of rational persuasion are *pro tanto* wrong, including in some ways that have not previously been remarked upon. Finally, we demonstrate that online settings possess a number of features that are especially conducive to wrongful rational persuasion.

Keywords Ethics · Persuasion · Rational Persuasion · Manipulation · Influence · Internet · Rationality

1 Introduction

Rational persuasion is often considered to be a morally innocuous kind of influence.[1] Indeed, it is the go-to contrast to manipulative, coercive and other problematic methods of influence (Dworkin, 1988, 154–6; Baron, 2003, 50; Greenspan, 2003, 162–3; Cave, 2007, 141–2). Authors who take rational persuasion to be unproblematic do not typically offer an account of it, or acknowledge that there are exceptions to its

[1] By 'influence', we just mean 'mental influence' and we take this to consist in intentionally changing another's attitudes. Many of the influences we discuss also intentionally alter the influencee's behaviour, and could thus aptly be described as behavioural influences, but we focus on mental influences in this article. Where influence on behaviour is salient, we focus on the corresponding intentions, desires, fears, and other attitudes that are closely tied to behaviour.

✉ Thomas Mitchell
thomas.mitchell@philosophy.ox.ac.uk

[1] Oxford Uehiro Centre for Practical Ethics, University of Oxford, Oxford, UK

[2] Jesus College, University of Oxford, Oxford, UK

being unproblematic. The thought is often that, so long as one limits oneself to influencing other people via rational persuasion, one's audience can still make up their own mind about what to believe or do, so their autonomy is respected and no moral wrong is committed. However, we think that this picture is not quite right. In this article, we argue that rational persuasion can be *pro tanto* morally wrong[2]—henceforth just 'wrongful'—in a wider range of circumstances than has previously been recognised and demonstrate some of the ways in which it can be wrongful. We then argue that this has important implications for the assessment of online communication: we show that online settings possess features that are especially conducive to these wrongful forms of rational persuasion.

We begin by setting out an account of rational persuasion, based on the idea that it involves influencing another by way of providing reasons. This account is intended to capture the uncontroversial, core instances of rational persuasion; we do not claim that it captures *all* instances. This means that, when we examine a case of rational persuasion, it should count as rational persuasion even by the lights of those who do not give an explicit definition. We then give a list of plausible explanations of why allegedly wrongful modes of influence, such as coercion and manipulation, are wrongful. We do not favour any one of these explanations over any other. Instead, we simply assume that each of these explanations succeeds in picking out a genuine 'wrong-making' feature of certain forms of influence—a feature possessed by some forms of influence and in virtue of which those forms of influence are wrongful. Next, we present some ways of influencing others that both fit the description of rational persuasion and bear at least one of these wrong-making features. Thus, we will have identified some forms of rational persuasion that are, on our assumptions, wrongful. To be clear, we do not argue that it is the *rationality* of the persuasion that makes it wrongful, but rather that being rational does not prevent rational persuasion from being wrongful. Finally, we show that online settings possess features that are especially conducive to these wrongful forms of rational persuasion. This is because they possess particular features that the offline world lacks which either make wrongful rational persuasion more likely to occur, likely to be worse when it does occur, or likely to affect larger numbers of people when it occurs.

This is a conclusion whose novelty can be seen from two perspectives. First, although it has been argued previously that rational persuasion is not always as innocent as is often assumed,[3] we will identify some varieties of wrongful rational persuasion that have not previously been explicitly recognised. Second, the application of this point to the online realm, on which we focus, has so far been underexplored.[4]

[2] We take an act to be *pro tanto* morally wrong just in case it is morally impermissible in the absence of a defeating consideration – that is, a stronger countervailing *pro tanto* reason. Thus, an act that is *pro tanto* wrong may be all-things-considered justified.

[3] See, for instance, Tsai (2014), Gorin (2014a; b), and McKenna (2020). Kenneth Einar Himma (2007) discusses the harms of information overload, which may sometimes result from instances of rational persuasion.

[4] Although the issue of the extent to which we are influenced online and the ethics thereof have been discussed previously, for instance in Specker Sullivan & Reiner (2021) and Sahebi & Formosa (2022), the issue of rational persuasion has not, we think, been sufficiently addressed.

2 What is Rational Persuasion?

Persuasion is, at minimum, a way of influencing another that does not depend on force or coercion. This is an imprecise characterisation, but our focus is on specifically *rational* persuasion, and, as it happens, we will not need a precise definition of persuasion in general.

What does 'rational' mean in this context? One thing that it cannot mean is 'indifferent'. When a decision, discourse, or thought-process is described as rational, there is sometimes an implication that we have no prior leanings in dealing with the matter at hand—we are simply even-handedly appraising or presenting the evidence before us without any commitment to arriving at one or other conclusion. However, persuasion always involves pushing one view over another. If you are talking to a friend about some subject in an entirely indifferent manner—that is, without the goal of getting her to accept any particular view on the subject—then, whatever you are doing, you are not persuading her. True, she may come to be persuaded in her views by your non-persuasive speech, but your behaviour is not persuasion in the sense relevant here. There is a difference between a persuasive act—what we call simply 'persuasion'—and an act that merely has the effect of persuading.

This does not mean that persuasion must always be partial or biased. Suppose that, on a given question, the evidence weighs much more heavily on one side than the other. A climate expert may seek to rationally persuade a sceptic of the reality of climate change by detailing the evidence both for and against it in an even-handed manner. They feel no need to overweight the reasons that support their view or underweight opposing reasons, since they are confident that the weight of reasons in favour of climate change will be sufficient to induce the sceptic to change her stance.

Rational persuasion, then, does not entail indifferent persuasion, which would be a contradiction in terms. So, how can we differentiate rational persuasion from persuasion more generally? Perhaps the simplest way is in terms of the means by which that outcome is to be achieved. For many complex actions, it is possible to distinguish and separately specify the intended outcome and the means via which it is produced. We have already touched in vague terms on what these are for persuasion: the intended outcome is to influence another's attitude(s) and the means must not include force or coercion. We propose that it is possible to demark *rational* persuasion by restricting the means to those that are 'rational': the giving of reasons in favour of a certain attitude.

This, then, is the account of rational persuasion that we will employ:

> *A* rationally persuades *B* to adopt attitude α if (i) *A* brings it about that *B* adopts α, and (ii) *A* does so only by giving *B* reasons for adopting α, and (iii) *B* adopts α on the basis of recognising (some of) the reasons given by *A*, and (iv) *A* intends each of (i)-(iii).

Note that this states a view on what is sufficient for rational persuasion. Although we think it plausible that (i), (ii), (iii), and (iv) are also necessary for rational persuasion, it is only the claim that they are jointly sufficient that we will need for our argument, so we will limit ourselves to that claim. Thus, we remain open to the possibility that there are forms of rational persuasion that are not captured by this account. Our goal is to offer an account that captures the core, uncontroversial instances of rational persuasion, and we think our account does this. This approach makes our account ecumenical to a wide variety of theories of rational persuasion, which is important given that not all authors writing on the subject present clear definitions. There may be disagreement over whether our conditions are necessary, but we think it would be hard to dispute that cases in which each of (i)-(iv) is fulfilled are cases of rational persuasion.

Condition (i) presents the aim of any persuasive act: to get the target to adopt the attitude in question. Condition (ii) presents the means by which that aim may be achieved. For instance, if *A* brings *B* to adopt α by hypnotising *B*, or by directly stimulating *B*'s brain so that they acquire the attitude, then *A* has not rationally persuaded *B* to adopt α (indeed, it is plausible that this does not count as persuasion at all). Or again, if *A* threatens or bribes *B* into adopting α, this would not count as rational persuasion. In one sense, threatening and bribing are instances of reason-giving, since the fact that one will get a reward for doing something or suffer a penalty for not doing so is itself a reason for doing it. However, we exclude from the category of 'reason-giving' cases in which the attitude α is produced by creating a practical (moral or prudential) reason to adopt the attitude. Thus, threatening, bribing and (dis)incentivizing do not meet condition (ii).[5]

The purpose of condition (iii) is to rule out cases of *B* adopting α for reasons unconnected to those that *A* has given. For instance, suppose that you want to persuade someone to get a new mobile phone. You give her a call and tell her how out-of-date and insecure her current phone is—how it may allow her personal data to be stolen—and urging her to get a new one. She does not care about privacy, though, so your reasons have little persuasive effect. However, as you are speaking, giving her your reasons, she notices that the sound quality of her phone is not quite up to scratch. She can understand you, but only by listening closely and effortfully, and she decides that she would be better off with a new phone. In this case, she has been brought to rationally desire a new phone because you have been giving her reasons to do so. But you did not rationally persuade her. It was not on the basis of her recognition of any of the reasons you presented that she changed her mind. It was the sound itself, rather than your reason-giving, that convinced her. It is to rule out such cases that we include condition (iii).[6]

[5] We allow that creating and presenting *epistemic* reasons is consistent with rational persuasion. When a known expert in meteorology declares that it will rain tomorrow, she creates and presents an epistemic reason—a piece of testimonial evidence—for the view that it will rain tomorrow, but it is nevertheless plausible that she engages in rational persuasion. We thank an anonymous reviewer for pressing us to consider cases of this sort.

[6] This is an example of the problem of deviant causal chains—see, for instance, Davidson (1980, 78–81), Peacocke (1979), and Stout (2010). We will not attempt to solve that problem here, but will assume that the 'right' kind of causation, whatever that might be, is encapsulated in the phrase, 'on the basis of'.

Finally, condition (iv) is there to mark the difference, mentioned above, between a persuasive act and an act that merely has the effect of persuading. If something you have said counts as a reason for another to adopt α and he consequently adopts it, you have not engaged in rational persuasion if you either didn't intend that he would adopt α, didn't intend to give him a reason to adopt it, or didn't intend that he would adopt it on the basis of the reason that you gave. There can be ambiguity over whether a particular act of influence was intentional. For instance, suppose that you addressed a large group of people with your message and a particular individual was persuaded, adopting the attitude in question on the basis of recognising some of the reasons you presented. You did not know that this individual was in the audience and the message was certainly not aimed specifically at them, so was your act of influence intentional? We take the view that if the intended target of persuasion is a group, then each individual within that group is also an intended target. It does not seem plausible for a speaker who has been trying to persuade a large audience to claim of an individual who becomes convinced by their reasons, 'I did not mean to influence them'.

Since our account refers to 'giving a reason', we should say something about what a reason is. We understand reasons in an objective, normative sense: a reason is a *fact* that *actually* counts in favour of someone adopting a given attitude. We remain neutral, however, on how exactly the 'counts in favour of' relation is best understood.[7] On this understanding, something that *appears* to count in favour of adopting an attitude, but in fact does not—either because it is not a fact, or does not actually count in favour of the thing—is not a reason. Thus, for example, getting someone to intend to drink the glass of deadly poison by telling them that it is just water does not count as giving them a reason to drink the poison.

There are, of course, different ways of presenting a reason. Perhaps the most obvious is to simply state the fact that is the reason. For example, to persuade a friend to give up smoking, one might simply state that 'smoking is bad for your health'. But one might be less explicit. For instance, you might try to get your friend to give up smoking by showing him an informative video about its effects and hoping that he pays attention and realises that he has reasons to stop smoking. Or, more subtly, you might show him a video of people smoking which contains fraction-of-a-second images of smoke-damaged lungs, in order to subliminally enforce the message that smoking is bad for the lungs. These are all ways of presenting genuine reasons—facts that count in favour of giving up smoking. However, we will take it that only the explicit statement of reasons counts as 'giving reasons' in our account. Modes of presentation that bypass the target's conscious attention, or even those that are presented in an implicit or indirect way, are ruled out. This makes for a restrictive account of rational persuasion, which will exclude some cases that perhaps intuitively count as rational persuasion, but this suits our dialectic purposes. We aim to show that, even when we limit ourselves to core, uncontroversial instances of

[7] For a range of diverging views on this, see, for example, Scanlon (1998, 17–22); Raz (1999, 15–6); Dancy (2000, 1–5); Parfit (2011, 31–8).

rational persuasion, it can be wrongful, including in ways that have not previously been noticed.

Having established a set of jointly sufficient conditions for rational persuasion, we will now consider what is wrong with wrongful interpersonal influence. When someone influences another in a wrongful way, what is it that makes it wrongful? In the subsequent sections we will turn to consider whether rational persuasion, of the kind that meets our jointly sufficient conditions, can be wrongful in these ways.

3 What is Wrong with Wrongful Influence?

There are various ways of influencing others that are *pro tanto* wrong, or at least, that often are so. Coercion, blackmail, exploitation, extortion, deception, domination, indoctrination, and manipulation are all terms that are often used to pick out particular forms of wrongful—or often wrongful—influence.

For our purposes, the boundaries of these different categories are not important. Nor is it important whether any of these types of influence is invariably wrongful, or only typically so. What does matter is what makes them wrongful when they are so. In what follows, we distinguish five different features that plausibly make an instance of influence wrongful. In the next section, we will identify some circumstances in which rational persuasion possesses one or more of these features. Note that we are here discussing *pro tanto* wrongs; some, maybe all, of these kinds of influence may be justified in the right circumstances. Ours is therefore a weaker claim than the claim that such influences are always, in all circumstances wrong, all things considered.

Perhaps the most obvious way in which influence can be wrongful is by interfering with the influencee's autonomy, either by diminishing or constraining a person's capacity to decide for themselves, or by interfering with attempts to exercise that capacity. Coercion and manipulation are both frequently presented as (often) wrongful for this reason (Cave, 2007, 136; Gorin, 2014b; Cohen, 2018, 486). These forms of influence are said to prevent the influencee from making their own decisions or coming to their own conclusions. This may be the case even if the influencee believes at the time that they are acting autonomously. For instance, if you buy a product because you are threatened with assault if you do not (coercion) or as a result of subliminal advertising techniques (plausibly manipulation), then it seems that the choice to buy the product, if it counts as a choice at all, did not come *from you*. The idea that interference with autonomy is at least a *pro tanto* moral wrong is commonly endorsed. John Stuart Mill (2015, 74) and Joel Feinberg (1986, 52–97) defend versions of this view. Meanwhile, George Tsai (2014, 87) argues that one of the wrong-making features of paternalism is the interference in another's 'sphere of agency'. Similarly, Sarah Raskoff (2022, 953) raises the concern that nudge techniques might interfere with what she calls 'formative autonomy'—one's capacity to choose what one values and, by extension, the kind of person one is—in the context of patients making medical decisions. Patricia Greenspan (2003, 163) argues that interference with rational autonomy is at least part of what is wrong with manipulation. In the sphere of technology ethics, it has been pointed out that technologically-enhanced nudges can have an impact on our autonomy and that, when this impact is negative, it is ethically concerning (Burr et al., 2018, 762–4, 767–8). The general idea behind all of these claims is, we think, that people have a legitimate interest in deciding for themselves, at least with respect to certain aspects of their lives, and that reducing or constraining their capacity to do so, or interfering with attempts to do so, is therefore *pro tanto* wrong.

A second wrong-making feature often ascribed to certain forms of wrongful influence—especially manipulation and deception—is that they intentionally or carelessly induce the influencee to make mistakes. The simplest way this can happen is by being induced to believe something false. But even if what one is directly led to believe is true, one might be misled in other ways. For instance, one may be given a false impression of the beliefs, values, and intentions of the speaker,[8] or be misled into assigning inappropriate weight or valence to certain reasons (Buss, 2005, 226–9). For example, one might, through clever rhetoric or the use of framing effects, present bad reasons as good, or weak reasons as strong, and thereby cause another to over- or underweight some reasons. Robert Noggle (2020, 249–50) argues that manipulation necessarily induces a mistake in the manipulee. He defines manipulation such that it always involves getting the target to either adopt an inappropriate mental state, or respond to a reason in a manner disproportionate to its actual weight.[9] And this feature of manipulation is, he claims, what explains why manipulation is wrongful (Noggle, 2020, 251).

A third wrong-making feature, often ascribed to manipulation, paternalism, and domination, is that of disrespecting the influencee by treating her as less than a moral and/or rational equal (e.g., Kant, 2017, 225–6; Buss, 2005, 229–30; Cholbi, 2017, 126–8; Tsai, 2014, 90).[10] According to Sarah Buss, for instance, the manipulator treats the victim 'as an (autonomous) object in his world, a character in his plot, rather than as someone with whom he shares the world, someone whose plot interacts with his own in ways he has not himself plotted … to treat her this way is incompatible with treating her as an equal, where the inequality at issue takes the form of an asymmetry, or lack of reciprocity, in one's interactions with her'. This kind of influence need not interfere with the autonomy of its target, but it still fails to treat the other person as befits their status as a moral and/or rational equal.[11] Rather,

[8] Moti Gorin (2014a, 58–9; 2014b, 91–2) gives good examples of this kind of influence. A politician is trying to get elected, so finds out what most of the electorate want. They then give arguments and make promises aligned with the desires of the electorate, though they do not care about those issues themselves, only about getting into office. Moreover, when they are elected, they keep all the promises that they have made in order to secure a second term for themselves. The electorate are thus not misled about whether the promises will be kept and some may be convinced by the politician's arguments. But they are misled about the politician's own values, beliefs, and intentions.

[9] Noggle is concerned with manipulation in particular, whereas our concern is with wrongful influence in general. Nevertheless, his view does seem applicable here.

[10] This is not to say that we must always treat others as though they are as good or as clever as ourselves—sometimes that will not be the case. But others ought to be treated as just as morally valuable as ourselves, and as being capable of responding to moral, theoretical, and practical reasons.

[11] On the other hand, it may be that interfering with autonomy of another always involves disrespecting them, and it may be that it is wrongful for precisely that reason. If so, this third wrong-making feature subsumes the first. We remain neutral on this.

they are treated as a complex, perhaps even autonomous, *object* that is to be used rather than engaged with on equal terms. In the online realm, Jongepier and Wieland, 2022, 156–75) argue that microtargeting is a way of using people as a mere means and can be wrong for that reason.

A fourth way in which influence is often thought to be wrongful is simply by causing harm to the influencee. Of course, many influences—including some manipulative and coercive ones—are intended to benefit the influencee. Consider paternalistic nudging, for example, towards healthier food choices or greater retirement savings, which is often characterised as manipulative (e.g., Sunstein, 2015, 444–6; Wilkinson, 2017, 258–9) but is intended to benefit the nudgee. However, many instances of coercion (Feinberg, 1990, 211) and manipulation (Buss, 2005, 231) do cause harm to the influencee, and are plausibly wrong for that reason. Carissa Véliz writes that we are often harmfully influenced by technology companies' use of our data (2021, 23–6). Christopher Burr and Jessica Morley warn against the potential for harm done as a result of over-using digital healthcare technologies, particularly in the realm of mental health (2020, 77–8).

Finally, influences are sometimes thought wrongful when the influencer lacks the standing to influence in the way that she does. Standing is a concept most commonly invoked in relation to blame, which (when expressed) we might understand as a type of influence intended to induce certain responses, such as guilt, the recognition that one has acted wrongly, or an apology. Blaming can be wrongful because the blamee is not blameworthy. But it can also be wrongful simply because the person doing the blaming lacks the standing to blame, for example, when they are complicit in the relevant wrongdoing, they regularly perform similar wrongful acts themselves, or what has happened is none of their business.[12] A soldier may be blameworthy for carrying out immoral orders, but the officer who gave those orders is in no position to do the blaming. A habitually dishonest person lacks the standing to blame a friend for inexcusably lying. You should not usually weigh in on an argument between a pair of spouses to whom you have no personal connection and start assigning blame, even if it is clear to you who is at fault. In such cases the blamer acts wrongfully not because the blame is undeserved but because, in blaming, she oversteps the limits imposed by her situation, role, or relationship to the blamee. This point can be expanded to apply to persuasion (and perhaps other forms of influence, too). It is not difficult to imagine scenarios in which one has good reasons to offer another reasons but refrains because it is 'not my place'. The soldier's role in relation to the officer, for instance, will often preclude them from trying to persuade their commander to take a different course of action. It is more appropriate to try to persuade one's spouse that they should take more exercise than to similarly persuade an acquaintance.[13]

To summarise, it is sometimes thought that interpersonal influence is wrongful when and because it:

[12] See, for instance, Radzik (2011, 582), Todd (2012), Watson (2015), and Snedegar (2023). What explains the phenomenon of standing and how it can be lost is a matter we leave aside.

[13] George Tsai (2014, 107–9) makes a similar point when discussing what can render the giving of reasons disrespectful.

- Interferes with the influencee's autonomy
- Induces the target to make a mistake
- Treats the target as less than a moral or rational equal (disrespect)
- Leads to harm
- Is exerted by someone who lacks the standing to influence in that way

This is not intended to be an exhaustive list of the wrong-making features of wrongful influence. Doubtless there are others. The point here is that all are features of some instances of influence that plausibly make them wrongful. In what follows we will simply assume that each of these putative wrong-making features is *in fact* a wrong-making feature; forms of influence that possess any of these features are *pro tanto* wrong. We now turn to consider whether and when rational persuasion might possess these wrong-making features.

4 Wrongful Rational Persuasion

4.1 Audience-tailoring

To begin, consider those who would induce certain beliefs, attitudes, and behaviours in significant sections of a population in order to get what they want. Politicians want people to vote for them; corporations wish to sell more of their products or services; activists may want to prompt or encourage a social movement. All may present themselves as honest and unbiased persuaders, whose only motive is to deliver the best for their audience. For the sake of simplicity, we will focus on the politician seeking election, but what follows will be applicable in other cases, too.

Claudia Mills (1995, 108) gives a pertinent example, which she refers to as the 'audience-tailored message':

> When [a] politician … is in a blue-collar neighbourhood, hard hit by the recession, he talks about what he will do to create jobs. He doesn't quite get around to announcing to the conservative voters there his support for gay rights. The latter message he saves for a rally with a gay and lesbian lobbying group, where he somehow fails to mention his proposed budget cuts in government-sponsored medical (AIDS) research; this he saves for an upcoming rally with Citizens Against the Deficit.

This is acceptable, according to Mills (1995, 108), since a politician cannot talk about every issue on which they have a policy at all campaign events and it is therefore reasonable to focus on that which the audience considers most important. However, this assessment overlooks a crucial factor. The politician is not merely focusing on what his audience considers most important, but on what will get him elected. The conservative voters may care about gay rights as much as they do job creation; the gay and lesbian lobbying group may care about medical research as much as they do gay rights. But the politician avoids those issues with those groups because he believes that it will lose him votes. The electorate may be deprived of the opportunity to decide on the basis of a complete picture and arguably thereby treated as less than morally equal; the audiences are arguably being used as mere tools to accomplish the politician's goal of being elected. They are also misled. It would be acceptable to prioritise the audience's priorities, given the limitations of time and audience attention spans. But the politician deliberately avoids very relevant topics with each group. The conservative voters and the gay and lesbian lobbying group may expect him to talk about what is most relevant to them, so may be misled into thinking that he has nothing of importance to say on gay rights and medical research budgets, respectively. Thus, by strategically omitting relevant points with different audiences, the audiences are both disrespected and plausibly induced into having mistaken attitudes.[14]

This kind of case, wherein a message is carefully matched to an audience to maximise its impact, does fit the proposed account of rational persuasion. The speaker brings their audience to endorse a certain opinion only by giving them reasons for doing so, the audience does so on the basis of the reasons given, and the speaker accomplishes all this intentionally. Yet it also plausibly exhibits two of the wrong-making features we have mentioned: it plausibly induces a mistake in the target and it plausibly fails to treat the target as a moral equal. When it does so, we have reason to think that this method of rational persuasion is wrongful.

4.2 Standing to Persuade

Another circumstance in which rational persuasion is often problematic is where the persuader lacks the standing to persuade. Even if one has something persuasive to say on a matter, there are occasions on which it is not one's place to do so. Suppose that Bertha is facing an important choice. She has recently discovered that she is pregnant and is uncertain about whether to have the pregnancy terminated. On the one hand, the pregnancy is unplanned and she does not consider herself ready to be a mother. On the other, she has always wanted to have children, even if not right now, and she feels some moral compunction about having an abortion.[15] Now suppose that Alice, an acquaintance of Bertha's, has heard about Bertha's situation and has a strong opinion on which option she should choose. Alice knows exactly which one she would pick if she were in Bertha's situation. What is more, she has some very good reasons to back up that opinion—reasons which, she feels sure, Bertha would find convincing if she were to explain them.[16]

Let us suppose that Bertha has not asked for Alice's advice. Depending on the circumstances, Alice may lack the standing to attempt to rationally persuade Bertha,

[14] For similar examples, see Moti Gorin (2014a, 58; 2014b, 91–2). The protagonists in these examples use rational arguments to get what they want, but what matters to them is the *effectiveness* of those arguments, not their rationality. For more on how one can mislead by not meeting ordinary audience expectations, see Grice (1975).

[15] This is based on an example given by Sarah Raskoff (2022, 954). Her concern is nudging, whereas ours is standing, but we have in common the point that extremely personal and important decisions should not usually be intruded on by others without invitation.

[16] To avoid distraction, we deliberately refrain from fleshing out Alice's position. Our concern here is with whether Alice should offer an opinion, not what that opinion is.

and may therefore act wrongfully by doing so. There are at least two reasons why this might be the case. Firstly, there is the matter of timing. George Tsai (2014, 94–5) considers the case of being too quick to offer advice. Bertha has only just been presented with a choice which, though highly important, is not urgent; she has time to research and consider it carefully. When Alice immediately offers her unsolicited advice, she oversteps a line. The right time to offer advice is after it has been solicited, or, at least, after the advisee has first had the chance to think about the matter for herself, and thereby to exercise her autonomy in a sphere in which she has authority. Conversely, one can also be too late to give advice. If Bertha has already made a firm decision, and especially if she has taken steps that commit her to that decision, such as making an appointment with an abortion provider or informing her partner that they are going to be a parent, it would not be appropriate for Alice to weigh in with her opinion.

Secondly, there is the question of Alice's relationship to Bertha. If Alice and Bertha are very close friends, then it may be appropriate for her to give an unsolicited opinion. Their relationship is such that Alice is likely to have Bertha's best interests at heart and to know what she values most deeply. It could even be that she knows that Bertha is prone to making certain kinds of mistakes and is keen to help her avoid doing something that she will regret—this is surely part of the role of a good friend. If, on the other hand, Alice is barely more than a stranger, it is simply not her place to offer insistent and unsolicited advice to Bertha.

With respect to both timing and relationship, it is hard to draw a definite line between having and lacking standing. At what precise point is it no longer too early to give advice, and when does it become too late? How close and of what kind must the relationship be for the persuader to have standing, and with respect to which issues? We will not address these questions here; the vagueness does not prevent them from being genuine considerations. It also matters *how* the rational persuasion is framed. If, in either case, Alice presents herself as merely making suggestions, while fully acknowledging that the choice is Bertha's alone, then her persuasive act is more likely to be appropriate even if her timing or relationship are not quite within the correct (vague) boundaries. But it would be wrong for her to rationally persuade Bertha in a particularly forthright or forceful manner, without acknowledging that it is ultimately up to her.

Rational persuasion is, we assume, wrongful when and for the reason that the persuader lacks the standing to persuade in the way she does. However, in such cases, the persuasion will also, we think, often possess one or more of the other wrong-making features that we have identified. Firstly, in the case of being too quick, it may interfere with autonomy. Alice's act of rational persuasion does not totally remove Bertha's autonomy, of course. Bertha could simply ignore what Alice says and decide on the basis of her own values and preferences. However, Alice's action plausibly makes it *more difficult* for her to do this. She reduces, without entirely eliminating, Bertha's ability to decide autonomously. The situation is in some ways akin to teaching someone about a philosophical idea. If the teacher has an opinion about that idea, it is wise to withhold it, at least to begin with, to allow the learner to consider the idea for themselves (Tsai, 2014, 99). Giving one's own view too quickly, especially on a matter that is ultimately up to someone else, can impede them in having a ‘purer, more direct engagement with the reasons most centrally relevant’ (Tsai, 2014, 96). Furthermore, Alice interferes with Bertha’s autonomy in the matter of *when* to make her deliberations. Perhaps she plans to think about it properly later, when the situation has sunk in and she has more time to devote to the matter, yet Alice pushes her to begin considering it before she wants to. Thus, a likely effect of Alice’s action is that Bertha’s choice will be less autonomous.

Secondly, there does seem to be a sense in which the target is treated as less than equal. Tsai suggests that intervening too quickly in another’s deliberative process may demonstrate disrespect for their ability to reason and come to the conclusion that most suits them (2014, 96–9). When Alice seeks to persuade Bertha too soon, she is implicitly suggesting that Bertha is incapable of working out what she should do and that Alice is capable of doing so for her. It is disrespectful to place oneself above another in this way.[17] While Tsai applies this point to persuading too soon, we think that it is even more disrespectful when the persuasive arguments are offered too late; it implies not merely that the other person is unlikely to make the right choice on their own, but that they in fact did not make the right choice, so the suggestion that they are less capable of making good decisions is even stronger.[18]

Thirdly, in cases where the persuader lacks the standing to persuade, there is often also a danger that the rational persuasion will lead to harm. Suppose that, having been persuaded by Alice’s reasoning, Bertha were to take the option that she has argued for. Only when it is too late does it dawn on her that she would have preferred the other. Suppose further that this is not merely a case of looking back and rethinking things with hindsight, having lived with her choice for some time. Rather, it is a case of realising that, had she really engaged with the matter for herself at the time, she would have known then. But Alice impeded her in doing so. Either because she was too quick, or she did not know Bertha well enough, she persuaded her into making a choice that was not right for her. Even if she never found out that she was on the ‘wrong’ path, perhaps never taking the time to seriously reflect on her past decisions, she has still been harmed, since her situation is worse according to her own values and preferences than it would have been without Alice’s intervention.[19] Rational persuasion is also prone to cause harm when it comes too late. For example, persuading someone, however rationally, that the decision they have just committed to is wrong may lead to unwarranted self-doubt and anxiety. Not only might Bertha be concerned that she has made a bad decision, but she may also begin to doubt her own powers to make good decisions about important and personal matters more generally.

It seems, then, that when one lacks the standing to persuade in the way that one does, one’s persuasion may also frequently be wrongful for other reasons. Tsai

[17] Similar points about mistrust and disrespect are made by Michael Cholbi (2017) in relation to paternalism.

[18] Tsai (2014, 106–7) does acknowledge that it can be disrespectful to persuade too late, but does not mention that this may be even worse than when it is done too early.

[19] This is in some ways similar to what Chang (2017) and Raskoff (2022, 952) call ‘drifting’ into an option: choosing it without first autonomously settling on a stable preference for it.

(2014) has already made a more restricted version of this point. He considers cases in which a person is too quick to offer rationally persuasive advice, and argues that this can express distrust and disrespect through the implicit suggestion that the other person is less capable than oneself of making good decisions, and can intrude on the other's deliberations in a way that interferes with autonomy in that decision (2014, 97). He also acknowledges that it can be wrong to rationally persuade too late or in the context of the wrong kind of relationship because doing so is disrespectful (2014, 106–9). We agree with Tsai, however our analysis of such cases suggests that more may be wrong with them than a display of disrespect. One may also lack *standing* to persuade because of issues with timing and relationship, cause the *harms* of a missed opportunity, reduce *autonomy* in the sphere of when to deliberate, and induce the *anxiety* of having one's decisions thrown into doubt.

4.3 Persuading into Belief in a Falsehood

A third kind of rational persuasion that is plausibly wrongful is persuading someone into believing something that is known, or at least believed, by the persuader to be false. This is perhaps the most intuitive and simple way in which rational persuasion can be wrongful. It is plausibly a kind of deception. As we will see, however, it need not involve any actual lies.

Suppose that Charles wants to know whether the shop round the corner is open. Daphne has recently walked past it and seen that, unusually for the day and time, it is closed. However, for her own purposes, she wishes Charles to believe that it is open. Rather than simply lying—perhaps she has a strong but twisted conscience that does not permit such direct methods of deception—she decides to rationally persuade him that it is open. She tells him a series of truths which he takes to count as evidence of the shop being open. Moreover, these are facts which really would count as evidence, if the shop were open. For example, Daphne tells Charles that the shop is normally open at this time of day; that the sign displayed in its window lists the current time as within its opening hours; that this is a weekday; that he has talked with several others today who frequent that shop and none has mentioned an unscheduled closure. However, she never explicitly asserts that the shop is open; she sticks to the truth, yet brings Charles to believe what is false. He adopts this belief on the basis of the evidence that Daphne provides him with (or reminds him of), and she brings this about intentionally, so she rationally persuades him according to the jointly sufficient conditions given earlier.

However, this case does have some of the wrong-making features identified above. Most obviously, it induces a mistaken attitude. The whole purpose of this strategy is to bring the other person to have a false belief. It might also interfere with autonomy. Acting autonomously, on some views, requires a more-or-less accurate picture of the situation. Thus, if Charles were to act on the belief that the shop round the corner is open—by going to visit it, say—his action may not count as fully autonomous. Daphne will have prevented him from acting fully autonomously by giving him a false impression. It can also lead him to harm if he depends on her word. At the very least, Daphne will have wasted Charles's time if he acts on the belief that she has persuaded him into. There may be further types of rational persuasion that are likely to possess one of the five wrong-making features that we have identified. These three, however—strategically selecting an audience, lacking standing to persuade, and persuading into falsehoods—will serve to demonstrate the point that, though rational persuasion is generally thought to be—and very often is—innocuous, there can be instances in which it is wrongful to rationally persuade. Furthermore, as we will see in the next section, these kinds of case are especially relevant to the assessment of digital technologies, since online contexts are highly conducive to these wrongful forms of rational persuasion.

5 Application to Online Contexts

In this section, we consider the three types of often-wrongful rational persuasion discussed above in turn, in each case highlighting how online contexts can facilitate wrongful rational persuasion of that type. We think that there are at least five features of online settings which either make wrongful rational persuasion more likely to occur than in typical offline settings, or likely to be worse when it does occur (because it bears the wrong-making features established above to a greater extent), or likely to affect larger numbers of people. These are as follows:

- Size: the amount of online content is immense and diverse, coming in many formats, covering many topics, and having many purposes. In particular, there is a vast amount of both information and misinformation.
- Connectivity: as of 2021, the Internet was accessible to some 63% of the world's population (The World Bank Group, 2023) and, in principle, each is made accessible to all others. One can potentially contact vast numbers of people through the online world.
- Speed: that content and those users can both be accessed extremely quickly – far more quickly than by using offline methods.
- Precision: it is possible, often using algorithms, to access precisely the content that one wants (e.g. search engines) and the audience that one wants (e.g. targeting on social media) (Zuboff, 2015).
- Disinhibition: perhaps due to anonymity, users are more willing to type in comments and messages that they would refrain from saying face-to-face (Stuart & Scott, 2021, 9).

There may be other features besides these that make online settings particularly conducive to wrongful rational persuasion and those listed here overlap with one another to some extent, so this should not be considered either a jointly exhaustive or mutually exclusive list of relevant features. However, all that is needed for our argument is that they are features of online settings and it will be shown below how each can facilitate wrongful rational persuasion. We leave aside the extent to which these features may be necessary or contingent features of the online world. Perhaps without some (combination) of them, online settings could not exist. Or maybe they could exist, but only in a radically different form. Or maybe it would be possible to remove one or more without completely overhauling the online world. We suspect that none could be removed without at least drastic change, but our argument does not depend on this.

In many of the cases we will discuss, questions could be asked about *who* exactly is doing the rational persuasion, and *who* is acting wrongfully. We will assume, in all of the cases that we discuss, that there is some individual or organisation who creates and/or posts a message online and who meets our conditions for rational persuasion. That is, the individual or organisation—henceforth sometimes 'the content provider'—intentionally brings it about that someone adopts some attitude α by giving that person reasons for adopting α, on the basis of which that person adopts α. And we will assume that, if anyone is *wrongfully* rationally persuading others, it is these content providers. However, in several of the cases we will describe, it could be argued that others—for example, the digital platform that distributes messages, or the collective of content providers—are also engaged in rational persuasion, and are also acting wrongfully. It might even be the case that 'software agents', such as targeted advertising and content-recommending algorithms, can themselves wrongfully influence users of online sites, not just the human agents who create the relevant content, despite the fact that such agents lack intentions or any other mental states (Keeling & Burr, 2022). We will not take a stance on this, though we think it is a potentially fruitful avenue. It has been suggested, for instance, that artificially intelligent agents can be manipulators and that they may not be able to influence us in non-manipulative ways (Klenk, 2020, 96–7; 2022, 101–2). Whether this is so, and whether it precludes such agents from engaging in (perhaps manipulative) rational persuasion should be pursued in future research.

5.1 Selecting an Audience

Let us turn to the first of the three types of often-wrongful rational persuasion: the strategic matching of audiences to messages in ways that omit relevant information. By ensuring that each audience receives the message most likely to have a persuasive effect, an individual or organisation can gain voters, customers, donors, or supporters, depending on the nature of the message and their aims in disseminating it.

Above, we discussed a case in which a politician selected his messages to suit the audience. Such targeting is common both online and offline. However, online contexts also facilitate a further type of targeting: selecting an audience to suit one's message. Using traditional, offline methods, it would be almost impossible to target the audience with any great precision. With modern digital technology, however, it is not only a possibility but a regular occurrence for a content provider to have control over who the audience is. Using sophisticated algorithms and extensive data, social media sites direct posts and advertisements to those most likely to respond to them. Politicians can thus more easily access their support bases; advertisers can find those more willing to buy their products; charitable organisations can target those most easily induced to make donations. Facebook, for example, has been found to be effective in matching adverts to those most susceptible to them. A 2017 study showed how psychological profiles of users can be built based on what they

'Like' and that more effective advertisements can then be shown to those users on the basis of their profiles (Matz et al., 2017). Or take a later study on the effects of microtargeted Facebook adverts on voter turnout in the 2018 Texas midterm elections. It was found that, although digital adverts had little effect in the aggregate (-0.04 percentage points compared to the control group), when certain salient issues were addressed, they had a significant effect on some voters. Specifically, advertisements related to abortion services increased the turnout of female voters by 1.86 percentage points in competitive congressional districts (Haenschen, 2023). The issue should not be overstated; these techniques may not (yet) be swinging elections in a democracy-endangering fashion. A recent study found that the efficacy of microtargeting was highly limited and dependent on context (Tappin et al., 2023, 6). Another study suggests that Internet advertising, while it can be effective, is not necessarily more cost-effective to the advertisers than more traditional forms like television advertising (Shaw et al., 2018, 372–3). There is not yet enough evidence to justify strong claims about the extent to which it influences voters, let alone the role it will play in future advertising and political campaigns as the technology advances (Fowler et al., 2020, 134–5). But it is clear that digital technology is able to match a message to the audience most likely to respond favourably to it. In these cases, the messages in question may or may not be *rationally* persuasive. But the technique does not essentially depend on lying, misleading, triggering biases, or any other morally dubious or non-rational method of influence; it merely relies on matching the right messages to the right audiences.

In precisely targeting an audience, one can do more than simply bring certain matters to their attention ('This is a product you might be interested in', 'This is a cause you may want to support', etc.). As the studies show, it is possible to persuade in the manner that a particular individual finds most convincing, appealing to various different types of evidence or practical reasons depending on what the target is most susceptible to. If this can be done for large numbers of people, it is easy to see how a significant impact can be had on a population. There are plenty of non-rational ways of persuading or influencing one's target audience, which may be troublesome for similar reasons. But the point relevant here is that, even if we restrict ourselves to consideration of only rational persuasion, we are still faced with the morally problematic features of matching audiences to messages discussed in §4.1 above: the audience are not treated as moral and/or rational equals and are misled about the intentions and beliefs of the speaker. Although it is the audience being selected on the basis of the message, not just the message being adjusted to suit the audience, it is still the case that recipients receive only part of the information; the persuader strategically omits some reasons from some audiences not on the basis of what is most relevant, but on the basis of what is likely to be most effective.

We here see two key features of online settings being brought to bear. One is connectivity; one can reach a far larger audience than is possible with offline campaigning methods. The other is algorithm-facilitated precision; the persuader can control who receives which persuasive messages. Even if it does not always work perfectly, this brings the politician, advertiser, or other persuading agent much closer to picking out all and only the audience that is most susceptible to each of their messages than they could without going online. The wrongs associated with audience-message matching are thus likely to occur more often and to more people online than offline because of key features of the online realm.

5.2 Online Standing

Next, consider the issue of persuading while lacking standing to do so. We saw that this can be morally problematic, not only due to the lack of standing, but also, for example, because persuading prematurely, too late, or within the context of an insufficiently close relationship can lead to a range of harms, can treat the target as less than an equal, and can interfere with autonomous decision-making. It is not difficult to see how these problems could arise online.

One way in which this can happen is in the context of parasocial relationships. These are unreciprocated emotional bonds sometimes formed with celebrities and media figures, in which one feels a certain closeness or intimacy with the other person, as if they were a personal friend, despite the fact that they usually do not even know of one's existence (Forster & Journeay, 2023, 714–5). These can form via traditional mass media such as television: you do not merely admire your favourite sports star, you feel as if you *know* them; you are not just entertained by the latest popular singer, you feel as if they are a part of your life. But they can form especially quickly and strongly with celebrities with a significant social media presence – a fan can easily feel as if they are getting a privileged insight into the celebrity's mind, a behind-the-scenes glimpse into their authentic private life, although this is unlikely to actually be the case (Hoffner & Bond, 2022).[20]

Now, consider Bertha's situation as she deliberates about her pregnancy. We mentioned before that Alice, in giving her advice, may lack standing to persuade because of either her timing or her relationship to Bertha. Now, imagine that Alice is not an acquaintance of Bertha's, but an Internet celebrity, an 'influencer' whom Bertha takes note of because, at least in the version of herself that she presents to her fans, she has experience in matters of both abortion and motherhood, and seems to talk of them with sensitivity, wisdom, and compassion. Furthermore, Bertha has formed a parasocial bond with Alice in the time that she has been following her various social media presences. However, she does not look for Alice's opinion on the subject; the parasocial relationship notwithstanding, she wants to be more rigorous in her approach than that. In fact, she does not seek anyone else's opinion for the time being, either in person or online, for she wants to spend time carefully reflecting on the matter for herself before taking advice.[21] Nevertheless, her online activity gives

[20] Do online 'influencers' have a special kind of standing towards those who voluntarily follow them, especially if they develop parasocial bonds? Given that it is a thoroughly one-way relationship, to the extent that the 'influencer' is unlikely to even know of the existence of an individual fan, we take it that whatever grants standing to close friends and family is lacking. However, we cannot give a definitive answer on whether this is always the case without delving into the details of what gives standing, which would take us too far off topic. In the case to be discussed below, we take it that Alice does not have standing to persuade Bertha.

[21] If Bertha were to start looking for information or reasons from either her friends or the Internet, then she is signalling that she is ready to start taking others' advice, so Alice might therefore not lack standing

her away; she is not specifically searching for topics related to abortion or childrearing, but the information gathered on her activities matches the profile of someone considering whether or not to terminate a pregnancy.[22] Accordingly, the relevant algorithms direct some of Alice's posts and videos to her social media newsfeed. In this content, Alice gives rationally persuasive messages to her audience about which choice would be better.

Bertha does not invite Alice's input. Alice nonetheless provides her with rationally persuasive messages to take one option rather than another. Given the parasocial relationship, Bertha is even more likely to take her advice to heart than if she was merely an acquaintance; this is someone whom she imagines to be her friend. However, she is in fact little more than a stranger and therefore lacks standing to an even greater extent than in the original case. Furthermore, since Alice is giving advice unprompted and does not know Bertha's situation, the timing is also even more likely to be inappropriate. Thus, the Internet-based parasocial relationship both exacerbates the wrong of persuading without standing and makes it more likely.[23]

The case of the parasocial relationship is one example, but online settings facilitate many other interactions involving rational persuasion without standing. They erode many of the social barriers that would otherwise exist between strangers, making it acceptable to interact on a more personal level (Stuart & Scott, 2021), including persuading others of certain points of view. Users often feel more in control and safer with online interactions than with offline ones, so are more disinhibited than they would otherwise be, especially those with higher levels of social anxiety (Scott et al., 2022, 298–9). Strangers online will always lack standing to persuade on intimate matters for want of the right sort of relationship, and the speed characteristic of online settings also encourages the giving of advice too soon. So, even without a parasocial bond, online settings lend themselves to rational persuasion without standing.

These issues are exacerbated by the fact that, on many social media platforms, the user has little control over what is recommended to them or the posts that they see. Content is presented without any regard to standing at all, but only according to what is deemed 'relevant' by the given algorithms. The only aim is to elicit the desired response from the user which is done based on the data gathered about them. Whether the person who made a particular message—an advertiser, for

Footnote 21 (continued)
on the basis of timing, although she plausibly would still lack standing on the basis of relationship. We thank an anonymous reviewer for making this point.

[22] This is an imagined but realistic example, based on a case of a teenager's pregnancy being revealed to her family after receiving coupons for childcare products, because a supermarket had inferred that she was pregnant from the data gathered on her activities (Duhigg 2012). It has also been shown that personal information such as religion, sexual orientation, and political affiliation can be inferred with a high degree of accuracy merely by analysing Facebook Likes (Kosinski et al. 2013). With much more data than this being gathered—on online searches, purchase history, social media activity, and so forth—both a pregnancy and the target's indecision about it could likely be revealed.

[23] Is condition (iv) of rational persuasion met if Alice does not know that Bertha is in her audience? Yes – as mentioned in Sect. 2, if a group is an intentional target of persuasion, then so is each individual within that group.

instance—has appropriate standing is simply not a consideration when it comes to who is shown what. If a rationally persuasive message relevant to some choice that a user is facing is shown to them, either in the form of an advertisement or a post from another user, this may be a problematic form of rational persuasion. Due to data gathering on her social media activity, Bertha can be very quickly shown a series of messages strongly suggesting, with reasons, that she pick one option over the other on an extremely personal matter. Even when these messages do not come from someone with whom she has formed a parasocial bond that makes her particularly susceptible to them, they would most likely come from people who do not know her well and for whom the decision is not their business. They might also come too late, after Bertha has made up her mind, which, as we argued earlier, can also be problematic.[24]

It therefore seems that online settings are not merely further fora in which persuasion without standing can take place. The fact that everyone can be connected to everyone else, the speed at which these connections can be made, and the lowering of social barriers that being online facilitates, means that the online realm exacerbates both the frequency and the severity of persuasion without standing.

5.3 Persuasion into Believing a Falsehood

We then have the issue of convincing another of a falsehood by appealing to truths that constitute evidence for it. This, again, is something that is made easier by the online world. Fake news and conspiracy theories are well-known hazards, but one need not delve into the wilder stories that spread online in order to find seriously misleading, and convincing, claims. It is possible to convince others of a falsehood by using only rational persuasion, particularly those who are not themselves experts in the relevant field.

Take climate change, for instance. Most of us are not climate scientists and cannot, without deferring to expert opinion, determine whether, how, and to what extent the climate is changing, whether this is caused by human activity, and what can be done to stop it or mitigate its effects. Suppose that someone wants to persuade others not to believe in human-caused climate change and to do this using only evidence. They go online and gather all the evidence that they can find, including testimony from the small number of well-qualified scientists who are climate sceptics, records of measurements and observations, and details of climate change in aeons

[24] There is a question, in cases like this, regarding whether the individuals generating or distributing the messages have the intentions required to count as engaging in rational persuasion. They may, for example, merely be expressing their views without any intention of changing others' mental states. There is a further question about whether they intend to or could foresee that they persuade *without standing*. If the lack of standing is neither intended nor foreseeable, some of the problems that we ascribed to persuading-without-standing above might not apply. For example, it is doubtful that one fails to treat another as an equal if one rationally persuades while unforeseeably lacking the standing to do so. However, at least in many cases, the message creators and/or distributors will, we think, count as rational persuaders, and in many of these it will be at least foreseeable for the persuader that they lack the standing to persuade. In these circumstances, we think that all of the problems that we identified in §4.2 above may arise.

past before humanity existed. This will, of course, be weak evidence when compared to the evidence for human-caused climate change, and there will be much less of it. Most climate scientists would be unconvinced and be capable of showing that the totality of evidence overwhelmingly points the other way. But given the size of the Internet, the sheer amount of content that it has to offer, it will be possible to find what seems to a layperson to be an impressive quantity of evidence relatively quickly. The climate-denier proceeds to disseminate their findings through various online means: they set up a website displaying the (highly selective) evidence in detail, write a series of blog posts about it, create videos in which they present their views, and post some of their favourite 'discoveries' on social media. They reach a large audience, many of whom become convinced that human-caused climate change is a myth, or at least is not well-supported by current evidence. The climate change denier has brought about their new attitudes solely by presenting evidence and the readers have been convinced by recognising the evidence as reasons to believe that human activity is not causing climate change, or at least be sceptical that it is. This is all intended by the climate denier, so counts as rational persuasion. True, they are very selective about the evidence that they present, but so long as they present nothing other than evidence, our account implies that they are rationally persuading their audience.[25]

It may be thought that this is an unrealistic example. Since only weak evidence can be given against climate change—or at least, only evidence that is significantly outweighed by countervailing evidence—surely hardly anyone would actually be convinced. In general, it may be possible to rationally persuade large numbers of people to believe a falsehood only on those odd occasions when the falsehood is better supported by evidence than the truth of the matter, or at least is not significantly outweighed. Therefore (one might suppose) the risk of rational persuasion of a falsehood online is much smaller than we have made it seem.

To this, it can be pointed out that it does not matter if the evidence is generally unconvincing. Even if it only convinces a small proportion of those who see it, the sheer number of people who can be exposed to it very quickly online means that this small proportion would still be a large number. Again, size and connectivity work to the advantage of the persuader. Moreover, it turns out that rational persuasion invoking weak evidence is often enough to neutralise the effect of countervailing strong evidence. A study carried out by Sander van der Linden and collaborators on how to combat misinformation about climate change showed that participants who were given what appeared to be evidence that significant numbers of scientists did not believe in human-caused climate change (a petition signed by 31,000 supposed scientists) adjusted their beliefs in the scientific consensus on climate change downwards accordingly. Perhaps more surprisingly, however, when this relatively weak evidence—31,000 seems like a large number, but is small compared to the

[25] This example is loosely based on the real-life case of Dan Peña, who has spread climate misinformation especially via TikTok videos (Silva & Ahmed 2023), although there are plenty of other similar cases. Our example is charitable to the climate change denier in that we think their claims are not usually based on evidence; the point is that, even if they were, it would still be ethically dubious.

total number of climate scientists—was presented alongside the information that 97% of climate experts did believe in human-caused climate change, the two cancelled each other out. There was no significant net change in either direction on what participants believed to be the level of scientific consensus on human-caused climate change (van der Linden et al., 2017).[26] This suggests that it does not take much to keep matters confused. If it suffices for some person's cause to maintain uncertainty on a given issue, then weak evidence is quite enough, even when the other side can produce strong evidence. Plausibly contributing to this effect is information overload: it is time-consuming and effortful for the recipients of large amounts of evidence—or any kind of propositional content—to sift through it and determine what is useful and what is not, what evidence is strong and what is weak (Himma, 2007, 271–2). Again, the size of the Internet plays a role – there is so much content available on almost any given topic (in this case, climate change), that it can be difficult to determine where the evidence points overall even if one is skilled at telling the difference between strong and weak evidence. Another contributing factor may be what has been called the *affective scaffolding* of online environments, especially social media. According to Steinert et al., (2022, 14), affective scaffolding 'designates those resources or elements of the environment that evoke, enable, support, enhance, regulate, sustain, and constrain affective experiences'. The kind of emotional atmosphere generated online potentially reduces users' tendency to think critically. That is not to say that any kind of emotion is opposed to critical thinking, but social media can create 'affective bubbles', wherein challenging the accepted narrative or being open to criticism of one's own views is discouraged, whereas vilifying perceived outsiders and shutting out their views is rewarded (Steinert et al., 2022, 18–22).[27] So, in the example given above, it may well be enough for the content provider to convince others that there is a substantial, even if uncertain, likelihood that climate change as a result of human activity is a myth.

As this example highlights, online settings provide an unparalleled opportunity for those who wish to mislead—including via rational persuasion—to reach a huge audience quickly. In particular, the size and speed of the Internet mean that large quantities of evidence, even if it is only weak evidence, can be gathered in a relatively short space of time. For almost any issue, it will be possible to find some form of evidence favouring one's preferred view. The precision of online tools like search engines also means that there is less need for effortfully searching through large numbers of sources—and those sources themselves are typically searchable, so one can even find small amounts of supporting evidence in a paper or article generally antithetical to one's views. Then, given online connectivity, such evidence can be disseminated to large numbers of people at once. The sheer quantity of content on the topic available online then makes it difficult for the recipients, if they are minded to check the sources, to determine the quality of the evidence. Intuitively, this counts

[26] Participants were asked, before and after being shown the 'evidence', what percentage of scientists they thought believed in human-caused climate change.

[27] We thank an anonymous reviewer for drawing our attention to the potential effects of both information overload and affective scaffolding.

as rational persuasion, and our account of rational persuasion can accommodate this verdict, since even weak evidence is a reason to believe.

6 Conclusion

The purpose of this paper has been to show that rational persuasion can sometimes be morally wrongful, including in some previously unnoticed ways, and that online settings are conducive to wrongful forms of rational persuasion. We began by developing a set of jointly sufficient conditions for rational persuasion:

> *A* rationally persuades *B* to adopt attitude α if (i) *A* brings it about that *B* adopts α, (ii) *A* does so only by giving *B* reasons for adopting α, (iii) *B* adopts α on the basis of (some of) the reasons given by *A*, and (iv) *A* intends each of (i)-(iii).

We then considered some of the wrong-making features that other forms of influence, such as manipulation and coercion, have been thought to have. We showed that certain kinds of influence that fitted the definition of rational persuasion also bore these putatively wrong-making features. Finally, we explained how online settings are conducive to these forms of persuasion. It is not merely that one can use online as well as traditional forms of communication to influence others in a wrongful fashion; online settings have features that either make wrongful rational persuasion more likely to occur than in typical offline settings, or worse when it does occur (because it bears the wrong-making features established above to a greater extent), or likely to affect larger numbers of people. These features are: size (vast quantity of online content); connectivity (access to enormous numbers of users); speed (relatively quick access to that content and those users); precision (efficiency in finding just the content or users that one wants); and disinhibition (users' willingness to post online what they would not say in person).

This argument does come with some important limitations. Firstly, the account of rational persuasion that we provide is not a full definition, but only a set of jointly sufficient conditions. It may well be that there are further kinds of rational persuasion (some of which may be wrongful) that this paper does not cover. Secondly—and more specifically—the fourth condition restricts us to considering only potential persuaders who have intentions. This means that we are not discussing the influence that algorithms themselves can have on users apart from matching them to (mis) information and arguments intentionally provided by other users.

Our discussion also, of course, leaves many important questions unanswered. Among these are the following. How seriously wrongful are the forms of wrongful rational persuasion that we have identified, and under what conditions might they, despite their wrongfulness, be permissible all things considered? How does the design of an online platform facilitate or limit the kinds of wrongs we have highlighted and can the online world be reformed to reduce the risk of those kinds of wrongs? If so, how could this be done? And could it be done while retaining those features of the online world that make it so useful? We leave these questions as topics for future research.

Acknowledgements This work was supported by a Consolidator Award from the European Research Council.

Author Contributions The first draft of the manuscript was written by TM. Both TM and TD commented on and edited previous versions of the manuscript. TM and TD both read and approved the final manuscript.

Funding This work was supported by a Consolidator Award from the European Research Council (project number 819757).

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Baron, M. (2003). ‘Manipulativeness’. *Proceedings and Addresses of the American Philosophical Association,* 77(2), 37-54.

Burr, C., Cristianini, N., & Ladyman, J. (2018). An analysis of the interaction between intelligent software agents and human users. *Minds and Machines, 28*(4), 735–774.

Burr, C., & Morley, J. (2020). Empowerment or engagement? Digital health technologies for mental healthcare. In C. Burr & S. Milano (Eds.), *The 2019 Yearbook of the Digital Ethics Lab* (pp. 67–88). Springer.

Buss, S. (2005). Valuing autonomy and respecting persons: manipulation, seduction, and the basis of moral constraints. *Ethics, 115*(2), 195–235.

Cave, E. (2007). What’s wrong with motive manipulation? *Ethical Theory and Moral Practice, 10*, 129–144.

Chang, R. (2017). Hard choices. *Journal of the American Philosophical Association, 3*(1), 1–21.

Cholbi, M. (2017). Paternalism and our rational powers. *Mind, 126*(501), 123–153.

Cohen, S. (2018). Manipulation and deception. *Australasian Journal of Philosophy, 96*(3), 483–497.

Dancy, J. (2000). *Practical reality*. Oxford University Press.

Davidson, D. (1980). *Essays on actions and events*. Oxford University Press.

Duhigg, C. (2012). ‘How companies learn your secrets’. Available at: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp. Accessed:17/02/2024

Dworkin, G. (1988). *The theory and practice of autonomy*. Cambridge University Press.

Feinberg, J. (1986). *The moral limits of the criminal law Volume 3: Harm to self*. Oxford University Press.

Feinberg, J. (1990). The *moral limits of the criminal law Volume 4: Harmless wrongdoing*. Oxford University Press.

Forster, R., & Journeay, J. (2023). Parasocial relationships and mental health. In H. Friedman & C. Markey (Eds.), *Encyclopedia of mental health* (3rd ed., pp. 714–719). Academic Press.

Fowler, E., Franz, M., & Ridout, T. (2020). Online political advertising in the United States. In N. Persily & J. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 111–138). Cambridge University Press.

Gorin, M. (2014a). Do manipulators always threaten rationality? *American Philosophical Quarterly, 51*(1), 51–61.

Gorin, M. (2014b). Towards a theory of interpersonal manipulation. In C. Coons & M. Weber (Eds.), *Manipulation: Theory and practice* (pp. 73–97). Oxford University Press.

Greenspan, P. (2003). The problem with manipulation. *American Philosophical Quarterly, 40*(2), 155–164.

Grice, H. P. (1975). Logic and conversation. *Syntax and Semantics: Speech Acts, 3*, 41–58.

Haenschen, K. (2023). The conditional effects of microtargeted facebook advertisements on voter turnout. *Political Behavior, 45*, 1661–1681.

Himma, K. E. (2007). The concept of information overload: a preliminary step in understanding the nature of a harmful information-related condition. *Ethics and Information Technology, 9*, 259–272.

Hoffner, C., & Bond, B. (2022). Parasocial relationships, social media, & well-being. *Current Opinion in Psychology, 45*, 101306.

Jongepier, F., & Wieland, J. W. (2022). Microtargeting people as a mere means. In F. Jongepier & M. Klenk (Eds.), *The philosophy of online manipulation* (pp. 156–179). Routledge.

Kant, I. (2017) *The metaphysics of morals*, trans. M. J. Gregor. Revised edition. Cambridge University Press.

Keeling, G., & Burr, C. (2022). Digital manipulation and mental integrity. In F. Jongepier & M. Klenk (Eds.), *The philosophy of online manipulation* (pp. 253–271). Routledge.

Klenk, M. (2020). Digital well-being and manipulation online. In C. Burr & L. Floridi (Eds.), *Ethics of digital well-being: A multidisciplinary approach* (pp. 81–100). Springer.

Klenk, M. (2022). (Online) manipulation: Sometimes hidden, always careless. *Review of Social Economy, 80*(1), 85–105.

Kosinski, M., Stillwell, D., & Graepel, T. (2013) 'Private traits are predictable from digital records of human behaviour'. *Proceedings of the National Academy of Sciences,* 110(15), 5802-5805.

Matz, S.C., Kosinski, M., Nave, G., & Stillwell, D.J. (2017) 'Psychological targeting as an effective approach to digital mass persuasion'. *Proceedings of the National Academy of Sciences,* 114(48), 12714-12719.

McKenna, R. (2020). Persuasion and epistemic paternalism. In G. Axtell & A. Bernal (Eds.), *Epistemic paternalism: Conceptions, justifications and implications* (pp. 91–106). Rowman & Littlefield International.

Mill, J. S. (2015). *On liberty, utilitarianism and other essays*. Oxford University Press.

Mills, C. (1995). Politics and manipulation. *Social Theory and Practice, 21*(1), 97–112.

Noggle, R. (2020). Pressure, trickery, and a unified account of manipulation. *American Philosophical Quarterly, 57*(3), 241–252.

Parfit, D. (2011). *On what matters* (Vol. One). Oxford University Press.

Peacocke, C. (1979). Deviant causal chains. *Midwest Studies in Philosophy, 4*(1), 123–155.

Radzik, L. (2011). On minding your own business: differentiating accountability relations within the moral community. *Social Theory and Practice, 37*(4), 574–598.

Raskoff, S. (2022). Nudges and hard choices. *Bioethics, 36*, 948–956.

Raz, J. (1999). *Engaging reason: on the theory of value and action*. Oxford University Press.

Sahebi, S., & Formosa, P. (2022). Social media and its negative impacts on autonomy. *Philosophy & Technology, 35*(3), 70.

Scanlon, T. M. (1998). *What we owe to each other*. Belknap Press of Harvard University Press.

Scott, R., Stuart, J., & Bareber, B. (2022). What predicts online disinhibition? Examining perceptions of protection and control online and the moderating role of social anxiety. *Cyberpsychology, Behavior, and Social Networking, 25*(5), 294–300.

Shaw, D., Blunt, C., & Seaborn, B. (2018). Testing overall and synergistic campaign effects in a partisan statewide election. *Political Research Quarterly, 71*(2), 361–379.

Silva, M., & Ahmed, M. (2023). 'The climate change-denying TikTok post that won't go away'. Available at: https://www.bbc.co.uk/news/technology-66023797. Accessed: 03/12/2023.

Snedegar, J. (2023). Explaining loss of standing to blame. *Journal of Moral Philosophy*. https://doi.org/10.1163/17455243-20234076

Specker Sullivan & Reiner. (2021). Digital wellness and persuasive technologies. *Philosophy & Technology, 34*(3), 413–424.

Steinert, S., Marin, L., & Roeser, S. (2022). Feeling and thinking on social media: Emotions, affective scaffolding, and critical thinking. *Inquiry*. https://doi.org/10.1080/0020174X.2022.2126148

Stout, R. (2010). Deviant causal chains. In T. O'Conner & C. Sandis (Eds.), *A companion to the philosophy of action* (pp. 159–165). Wiley-Blackwell.

Stuart, J., & Scott, R. (2021). The measure of online disinhibition (MOD): Assessing perceptions of reductions in restraint in the online environment. *Computers in Human Behavior, 114*, 106534.

Sunstein, C. (2015). The ethics of nudging. *Yale Journal on Regulation, 32*(2), 413–450.

Tappin, B., Wittenberg, C., Hewitt, L., Berinsky, A., & Rand, D. (2023) 'Quantifying the potential persuasive returns to political microtargeting'. *Proceedings of the National Academy of Sciences,* 120(25), 1-10.

The World Bank Group. (2023). 'Individuals using the Internet (% of population)'. Available at: https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2022&start=1960&view=chart&year=2021. Accessed: 01/12/2023.

Todd, P. (2012). Manipulation and moral standing: An argument for incompatibilism. *Philosophers' Imprint, 12*(7), 1–18.

Tsai, G. (2014). Rational persuasion as paternalism. *Philosophy & Public Affairs, 42*(1), 78–112.

Van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges, 1*(2), 1600008.

Véliz, C. (2021). *Privacy is power: Why and how you should take back control of your data*. Corgi Books.

Watson, G. (2015). A moral predicament in the criminal law. *Inquiry, 58*(2), 168–188.

Wilkinson, T. M. (2017). Counter-manipulation and health promotion. *Public Health Ethics, 10*(3), 257–266.

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology, 30*, 75–89.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.