*Journal of Ethics and Social Philosophy*
VOL. 31, NO. 6 · OCTOBER 2026
*https://doi.org/10.26556/jesp.v31i6.6271*
© 2026 Authors

# DISTINGUISHING PERSUASION FROM MANIPULATION

*Thomas Mitchell and Thomas Douglas*

IT IS OFTEN DIFFICULT to make a clear distinction between different methods of influence. When does an offer of rescue become a threat of abandonment? Where do we draw the line between bargaining and blackmail? At what point does a disincentive become coercive? These distinctions matter because there are *prima facie* moral differences between these various ways of influencing others. Being able to tell them apart plausibly helps us distinguish morally permissible and morally impermissible methods of influence.

One such distinction is that between persuasion and manipulation. It is usually considered morally permissible to persuade someone to believe or do certain things; the persuader exerts an influence but ultimately leaves it up to the persuadee. Conversely, it is typically considered presumptively impermissible to manipulate others, though the available explanations for *why* it is so are many and various; some appeal to the way in which manipulation interferes with the manipulee's autonomy, others to the harm it (typically) causes, others to the negligence or disrespect it expresses or constitutes, and still others to the mental mistakes that it induces.[1] Whatever the correct explanation, it is noteworthy that calling an act "persuasive" does not carry the same moral approbation as calling an act "manipulative." The various explanations just listed for the wrongfulness of manipulation are all based on accounts of what it is to manipulate; if an act of manipulation is wrong, then its wrongfulness is at least partly grounded in its manipulativeness. Conversely, while there can be cases of immoral persuasion, these are immoral for reasons other than their persuasiveness.[2]

1 See, respectively, Cave, "What's Wrong with Motive Manipulation?" 138–39; Buss, "Valuing Autonomy and Respecting Persons," 230–35; Klenk, "Interpersonal Manipulation"; and Noggle, "Pressure, Trickery, and a Unified Account of Manipulation," 249–50.

2 For instance, George Tsai argues that rational persuasion can be paternalistic and immoral on those grounds ("Rational Persuasion as Paternalism"). We have elsewhere presented several ways in which rationally persuading someone can be wrongful (Mitchell and Douglas, "Wrongful Rational Persuasion Online," 9–14).

It may be important to be able to tell the two apart, but is it difficult? There are plenty of clear-cut cases but also some that are unclear. For instance, suppose that you tell a friend that if they continue with a certain way of life, you will withdraw your affection from them, and the friendship will be over. On the one hand, it may seem manipulative to exploit their natural desire to be liked and to have friends in order to have them behave in a way that you condone. On the other, you could be informing them of where you regard the limit to be when it comes to your friendships, in the hope of thereby persuading them to do (what you take to be) the right thing. Alternatively, suppose that you persistently tell your friend to change their ways and repeatedly tell them why you think they are doing something wrong. Are you manipulatively nagging and pressuring them into doing what you think they should do, or are you merely persevering with your persuasive efforts? Or again, you might try to appeal to their moral emotions by telling them, "I thought you were better than this" or some such thing. Is this emotional manipulation, or are you persuasively appealing to their conscience or sense of decency? Whether cases like these count as manipulation or persuasion requires closer examination.

So too does the moral significance of the distinction. There are some cases of purported manipulation that appear quite innocuous. For instance, dressing smartly and smiling in order to give a good first impression or starting a discussion on a topic that bores you but interests your interlocutor whose friendship you wish to cultivate—if these are cases of manipulation that are not even presumptively wrong, then is manipulation a moral category worth delineating? Two things can be said to this. First, if these are instances of manipulation, they are only slightly manipulative. We therefore should not expect them to be seriously morally problematic even if manipulation as a general category is morally problematic. Flicking someone's ear is a genuine instance of causing pain, and causing pain is clearly a moral category worth delineating; that does not mean that we have to afford a central role to ear-flicking when giving accounts of the nature and ethics of pain. The more pain is caused, the more wrong the act is likely to be; similarly, the more manipulative an act is, the more wrong it is likely to be; and the more wrong a kind of act is, the more consideration to be afforded it in discussions of ethics. Second, even if it is granted that such minor manipulations are not wrong at all rather than being very small wrongs, it is still the case that paradigmatic manipulative behaviors are presumptively wrong. There may be exceptions or hard-to-categorize cases, but that does not mean that manipulation considered in general is not a concept with moral significance or is not worth examining. If anything, the fact that it is hard to draw clear boundaries between what is and is not manipulation and between what is and is not presumptively wrong counts in favor of closely analyzing the concept.

What does seem clear is that paradigm cases of persuasion are presumptively permissible, whereas paradigm cases of manipulation are not.

The purpose of this article is to develop an account of persuasion that marks a conceptual distinction from manipulation. This also serves as a guide for a moral distinction, although, as mentioned, there may be instances of manipulation that are not even presumptively wrong. Furthermore, the concept of persuasion stands in need of analysis, even apart from its border with manipulation. It is something that we spend a great deal of effort and time in doing, is a common reason for communicating with others, and the need to do it can motivate us to better interrogate our own views.[3]

Despite this, persuasion has received comparatively little philosophical attention. When it is mentioned, it is typically to draw a contrast with some other ethically important concept. There are two main ways in which this is done. The first is to consider specifically *rational* persuasion and highlight the apparent contrast between it and some morally dubious method of influence, such as manipulation, deception, or paternalism, occasionally with the aim of showing that there is not such a strong contrast after all.[4] The second is to contrast persuading with merely informing, a supposedly more innocuous form of influence.[5] This time, it is persuasion that is morally dubious, typically on the grounds that it can interfere with others' autonomy, whereas just providing information leaves the matter entirely up to the other person and even empowers them to make a choice that is more fully their own, especially in the context of obtaining consent. For instance, it might be thought that a doctor should not try to persuade a patient to agree to a certain treatment but only provide them with the relevant information. Thus, persuasion is often either treated as the "good" kind of influence, contrasted with the "bad" kind of influence that the author is really interested in (manipulation, coercion, etc.), or it is treated as the "bad" kind of influence, contrasted with the "good" kind of influence that the author wishes to highlight (informing).

What unites these discussions is that they do not comprehensively analyze the concept of persuasion itself. At most, a brief definition is provided. The concept of persuasion is taken as intuitively understood for the purposes of the author's argument—to show that rational persuasion can be paternalistic, as a

3 It has even been theorized that the need to persuade others of our own view is the evolutionary explanation for the development of human reasoning skills, complete with its flaws and biases. See Mercier and Sperber, "Why Do Humans Reason?"

4 See, for instance, Tsai, "Rational Persuasion as Paternalism"; Gorin, "Do Manipulators Always Threaten Rationality?" 51; and Cohen, "Manipulation and Deception," 486.

5 See, for instance, Crisp, "Persuasive Advertising, Autonomy, and the Creation of Desire," 413; and Brown, "'Sneaky' Persuasion in Public Health Risk Communication," 6.

stepping stone to offering an account of manipulation, or to show that certain kinds of advertising are immoral, for example.[6]

An exception is Keith Dowding, who does offer an account of persuasion.[7] Dowding takes persuasion to be changing another's beliefs by speech acts and investigates the question of when persuasion so understood is neither coercive nor manipulative. His account of noncoercive, nonmanipulative persuasion, which for him is a subtype of persuasion, bears similarities to the account of persuasion we propose below. However, as will be explained, we take our account to have advantages over his.

Perhaps some cases of persuasion are manipulative, and some cases of manipulation are persuasive. If so, then the morally important distinction is that between nonmanipulative persuasion, on the one hand, and manipulation—persuasive or nonpersuasive—on the other. Our aim is therefore disjunctive: if there is no overlap between persuasion and manipulation, then our account of persuasion should be read as clearly marking that distinction; if there is an overlap, our account should be read as distinguishing nonmanipulative persuasion from manipulation. It is thus ecumenical to whichever view the reader favors. Hereafter, references to persuasion should be taken to mean specifically nonmanipulative persuasion.

In order to construct our account, we need an account of manipulation to define persuasion against. In the next section, we introduce and develop our account, which is subsequently used throughout this article. It is a version of Robert Noggle's "mistake account."[8] Noggle gives six paradigm examples of manipulative techniques:[9]

1. *Nagging*: Pestering the target about an issue without providing any new reasons, with the aim of annoying them so much that they give in
2. *Playing on Emotions*: Instilling strong emotions such as fear, guilt, or resentment in the target so that they will decide to take a particular course of action
3. *Exploiting Reciprocity*: Doing small favors for the target so that they feel like they "owe" the manipulator and will therefore be more compliant

6 See, respectively, Tsai, "Rational Persuasion as Paternalism"; Cohen, "Manipulation and Deception"; and Crisp, "Persuasive Advertising, Autonomy, and the Creation of Desire."

7 Dowding, "Power and Persuasion."

8 Noggle, "Pressure, Trickery, and a Unified Account of Manipulation" and *Manipulation*, 102–19.

9 In his more recent work, Noggle extends this list to thirteen examples of manipulative techniques (*Manipulation*, 13–36). We keep to these six for simplicity; further examples would not change our argument.

4. *Gaslighting*: Undermining the target's confidence in their own judgment and perception so that they will be susceptible to the manipulator's suggestions
5. *Emotional Blackmail*: Threatening to withdraw affection, respect, or friendship if the target does not comply
6. *Selective Attention*: Keeping the target's attention on the positives of the desired behavior and off the negatives.[10]

We also take these to be paradigm instances of manipulative influence, though we do not claim that they are the only methods of manipulation.[11] Note that they do not include the kinds of science fiction cases often found in the literature on free will and moral responsibility involving brainwashing, hypnosis, evil neuroscientists, mind-control drugs, and computer chips implanted in brains, even though these are sometimes characterized as paradigm cases of manipulation and may, we allow, fall within the scope of the concept as defined by common usage.[12] This is not because we consider these too unrealistic; a thought experiment need not be realistic to be philosophically informative. Rather, they are not the kind of case with which we are concerned. Manipulation is often thought to be interesting and controversial in the ethics of influence in that, unlike an agent who is simply forced to do something, the manipulated agent still makes their own choices based on their perceived circumstances. There is thus an important sense in which manipulees retain their agency. Indeed, manipulators depend on this. For instance, when playing on emotions, a manipulator does not literally *force* their target to do something but creates circumstances in which the target is likely to *decide* to do it. It is this characteristic of manipulation—that it appears to simultaneously undermine and depend upon the target's agency—that makes it so difficult to identify, both conceptually and practically. In the "manipulation" cases involving brainwashing, hypnosis, and so on, the target's agency is entirely bypassed so they are not paradigm cases of manipulation in the sense that many ethicists of influence are primarily interested in. Indeed, we take them to be not cases of manipulation at all but rather severe cases of what Noggle calls *meddling*

10 Noggle, "Pressure, Trickery, and a Unified Account of Manipulation," 241–42.

11 We do not take there to be any difference between "manipulative influence" and manipulation. The former term is occasionally used to more clearly contrast manipulation with nonmanipulative influence.

12 See, for example, Frankfurt, "Alternate Possibilities and Moral Responsibility," 835; Pereboom, *Living Without Free Will*, 112–14; Mele, *Autonomous Agents*, 145–46; and Fischer and Ravizza, *Responsibility and Control*, 35.

or what Thomas Douglas elsewhere calls *mental interference*.[13] They are more akin to force since, although they operate in the mind, the target's own preferences, desires, beliefs, and other mental attitudes are entirely overridden, just as effective physical force overrides the target's physical actions. The kind of manipulation we are interested in, by contrast, makes use of the target's mental attitudes and processes for the manipulator's own ends.[14]

In the next section, we elucidate and develop Noggle's account. Based on this, we go on in the following section to show what it means to nonmanipulatively influence someone. The remaining section is dedicated to persuasion itself: defining it against both manipulation and other forms of nonmanipulative influence, offering an explicit account and enumerating the advantages of this account.

## 1. MANIPULATION

Taking inspiration from Noggle, we take manipulation to be inducing a mistake, where a mistake is a faulty mental attitude.[15] Below, we present and defend our account before showing why we consider it to be an improvement over Noggle's own view.

We do not claim that our account captures all forms of influence that common usage classifies as manipulation—that is, we allow that it may be somewhat revisionary. As mentioned earlier, common usage may classify as cases of manipulation some instances of, for example, brainwashing or electrical brain stimulation that our own account does not capture. As also mentioned earlier, if there is an overlap between persuasion and manipulation, then the morally important distinction is that between nonmanipulative persuasion and manipulation (rather than nonpersuasive manipulation and persuasion). Morally, manipulative persuasion should be considered manipulation, so that is where to draw the line. This is a matter of establishing a philosophically useful taxonomy rather than reflecting common usage. However, we do consider it to cover most of the cases commonly described as manipulation, including the paradigm examples mentioned above. Our account is as follows:

> *A* manipulatively influences *B* to adopt a mental attitude *a* if and only if (1) *A* intentionally brings it about that *B* adopts *a* and (2) *A* does so

13 Noggle, *Manipulation*, 207–8; and Douglas, *Protecting Minds*.

14 Our thanks to an anonymous reviewer for raising a case of this kind as a potential counterexample to the account we propose below.

15 Noggle, "Pressure, Trickery, and a Unified Account of Manipulation" and *Manipulation*.

while believing *a* to be a faulty attitude for *B*, where faultiness is understood as *a* being overall opposed by *B*'s reasons.[16]

This account is a variant of Noggle's view. It differs most significantly from his view in clarifying what a mistake is: a faulty attitude, where faultiness is overall opposition to the weight of one's reasons. Below, we explain the account further and in so doing show how it improves on Noggle's view while staying close to the spirit of it.

The first thing to note about this account is that it is concerned with attitudes rather than with behaviors. It might therefore seem limited in scope; one can manipulate others into doing (or not doing) things, not just into having beliefs, desires, or other mental attitudes. However, we believe that if one is manipulated into behaving a certain way, then one must have been manipulated into having some relevant desire, preference, or intention. This is because it is characteristic of manipulation that the target in some sense chooses how to act. Precisely what that sense is, given that there is also a sense in which the manipulator chooses how the target acts, is a primary source of the controversy over what manipulation is. But what is certainly not manipulation is getting the target to do something without their ever intending, desiring, or having any other behavior-inducing attitude to do so. That is simply forcing them. On the other hand, if someone has been manipulated into having a behavior-inducing mental attitude, such as a desire or intention, but then fails to execute the corresponding behavior, then they have still been manipulated. For example, suppose that a friend of yours is planning an adventurous holiday or road trip, but you want them to stay behind. You employ manipulative techniques like nagging and guilt-tripping (a form of playing on emotions) to get them to change their mind. Suppose that your efforts—constantly playing on their mind throughout their planning and packing—eventually succeed, but only at the last minute, when your friend is already driving to their destination. They decide to turn back but get confused about the directions and accidentally end up where they were going in the first place. It is not a failure of the manipulation if the target has adopted the relevant attitude but does not manage to follow through with the relevant action.

The second condition states not that *a* is a faulty attitude but that the manipulator believes it to be faulty. Well-meaning but bad advice is not manipulative.

16 The belief need not be occurrent or even conscious, but it must be related to the intentional bringing about of the attitude in such a way that *A* intentionally brings about a faulty attitude. (More on this below.) We could simplify our account to "*A* manipulatively influences *B* iff *A* intentionally brings it about that *B* adopts a faulty attitude," but we separate it into two conditions for ease of analysis later.

Nor is honestly but mistakenly claiming that another has misremembered some detail. This indicates that the influencer needs to believe that the attitude they are seeking to produce is a faulty one. In a similar vein, consider the reverse situation. An influencer tries to give bad advice or sow doubt in the target about their own memories, but it turns out that it is advice that the target really ought to follow or that the target really does misremember the relevant details. Such cases are most intuitively categorized as genuine though certainly nontypical cases of manipulation. Manipulation is certainly attempted, and it does not seem right to call it a failed attempt if they bring it about that the target adopts the attitude. What matters to the manipulator, after all, is getting the target to adopt that attitude. If the attitude turns out not to be faulty, that may be no harm to the manipulator's purpose. Therefore, what matters for manipulation is what the manipulator believes to be faulty, not what actually is faulty.[17]

As also implied by the second condition, we take it that a faulty attitude is one that is overall contrary to one's normative reasons—henceforth just *reasons*. There might be some reasons genuinely in favor of the attitude, but these are outweighed by opposing reasons. It is not our purpose here to give a precise definition of a reason, but we should say something about what it means to have a reason. In particular, we should clarify whether a reason, in this context, should be considered objectively or subjectively. If someone has a desire to drink a gin and tonic and believes that the liquid in the glass in front of them is gin and tonic, then they have reason to drink the liquid in the glass subjectively, from their own perspective. If, however, the liquid is really petrol, then they do not have reason to drink it in the objective sense, whatever they believe about it.[18] In referring to *B*'s reasons in the account above, should we use the objective or the subjective sense of the word?

We think that it is the objective interpretation that is relevant in the current context. Manipulation is concerned with mistakes, which fit more readily with an objective view of reasons than a subjective one. If, in the above example, the person who wants a gin and tonic has been tricked into thinking that the petrol in the glass is really gin and tonic and therefore desires to drink the contents of the glass, then it is intuitive that they have been manipulated into desiring the contents of the glass. But a subjective view of reasons does not capture this. On

17 This partly aligns with what Noggle says of this matter. He draws a distinction between the target *being manipulated* and the influencer *acting manipulatively*. In cases like those just described, Noggle holds that the influencer has acted manipulatively, but the target has not been manipulated; being manipulated requires making an *actual* mistake rather than making what either the influencer or the target believes to be a mistake. See Noggle, *Manipulation*, 89–93, 107.

18 This example is based on Williams, "Internal and External Reasons," 102.

a subjective view, their desire is still in accordance with their reasons, even if it is based on a false belief.[19] On an objective view, however, there is a clear mistake, for whether they believe it or not, the fact that there is petrol and not gin and tonic in the glass means that they do not have reason to drink from it. Furthermore, when a manipulator intentionally induces a mistake, they need not be trying to get the target to act against what the target takes to be their reasons. Indeed, one way of manipulating someone into doing something is to induce a change in their beliefs or desires so that their subjective reasons support that course of action. So we take reasons to be objective. More specifically, taking our cue from Joseph Raz and T. M. Scanlon, we take a reason for an attitude to be a fact that counts in favor of that attitude.[20] Again, it is not our purpose to give a full account of reasons, so we do not attempt to analyze the *counts in favor of* relation; for our purposes, an intuitive understanding suffices.[21] What matters for manipulation, then, is what the influencer believes to be their target's (objective) reasons, understood as facts counting in favor of attitudes, not what the target takes their reasons to be or what their reasons actually are.[22]

What do we mean by an attitude that is *overall* opposed by a target's reasons? Taking into consideration the reasons both for and against it, the weight of those reasons is against adopting the attitude. Adopting an attitude is overall *supported* by one's reasons just in case the total weight of reasons both for and against it favors adopting the attitude.

It is an important part of our view not only that the manipulator believes the attitude to be faulty and induces that attitude in their target but that these

19 Of course, the false belief is mistaken. But on a subjective view, even a false belief can be a reason.

20 Raz, *Practical Reason and Norms*, 18–20; and Scanlon, *What We Owe To Each Other*, 17.

21 Alex Kaiserman neatly captures the kind of thing we mean: "That apple crumble is tasty is a reason for me to eat some; that it is unhealthy is a reason for me not to eat some; and that it is typically baked at 180°C is neither. . . . Note that reasons are not desires, beliefs, or any other kind of psychological state, on this view; nor, in general, are they facts about such things" ("Reasons-Sensitivity and Degrees of Free Will," 689).

22 Analyzing concepts within the scope of a belief can be problematic, for the subject holding the belief may not agree with or have even thought about the analysis. Just as someone might believe "the morning star is a body illuminated by the sun" without believing "the evening star is a body illuminated by the sun," since they do not believe that the morning and evening stars are identical (Frege, "Sense and Reference," 214–15), so an influencer might believe "this attitude is faulty" or "this attitude is a mistake" without believing "this attitude is contrary to the objective reasons of my target." We think that the main arguments of this article holds even if the content of the influencer's belief is left unanalyzed so long as appropriate substitutions are made throughout. We provide the analysis for the sake of precision and to allow for meaningful discussion of faulty attitudes both within and without the scope of a belief.

factors are connected in the manipulator's mind in such a way that the fault is intended. The manipulator does not merely induce an attitude and also believe the attitude to be faulty; they induce the attitude *while* believing the attitude to be faulty. We wish to avoid classifying as cases of manipulation cases in which the fault is a foreseen but unintended consequence, as well as those in which the relevant mental states are not "put together."[23]

As an example of the former, suppose that you have a friend whose romantic partner is mistreating them, but your friend is blind to it, often making excuses for their partner's behavior or blaming themselves. Any time you try to raise the issue of their partner's mistreatment, they deflect it, change the subject, or simply ignore you. You therefore take a less direct approach by encouraging them to critically reflect on the value of all their various relationships—not just with their romantic partner but also with their friends, family members, and colleagues—and suggesting criteria for said evaluation. You intend that this exercise will result in them "raising the bar" for their relationships and ultimately realizing that the relationship with their partner is not worth maintaining, for it would not meet the new standard that your friend sets. However, you also know that a general raising of the standard for relationships will catch up some friendships that are actually good for your friend—perhaps including your own. Nevertheless, you think that this is a price worth paying if it means your friend can escape a potentially damaging romantic relationship before they are more seriously harmed. As it happens, your friend does indeed end not only their bad relationship with their partner but also several good friendships. On our view, you have not manipulated your friend into ending the good friendships. You did not intend that they should do so even though you intentionally induced the relevant attitude (of critical reflection towards relationships) and foresaw the negative consequences.

As an example of cases in which the relevant mental states are not "put together," suppose that you want a friend to leave you alone for a while but do not wish to hurt their feelings by letting them know that they are bothering you. You therefore ask them to perform some useless task to keep them occupied, thereby intentionally inducing in them a desire to perform that task. Since the task is useless, the desire is faulty—it is opposed by their objective reasons—so you have induced a faulty attitude in them. Moreover, you know and therefore believe that the desire is faulty, although that knowledge is not at the forefront of your mind when you make your request and is certainly not part of your motivation for making it. You are just trying to get them to leave you alone without hurting their feelings. So although you intentionally induce

23 Our thanks to an associate editor for raising this point.

an attitude (the desire to perform the task), and you believe that said attitude is faulty, you do not intentionally induce a fault, for your intention to induce the attitude and your belief that the attitude is faulty were not "put together" in your thinking about your action. We do not think that a case such as this should count as manipulation. It is possible, then, that someone might intentionally induce an attitude in another and simultaneously believe that attitude to be faulty, without intentionally inducing a fault in them, whether because they have not made the relevant inference or because inducing a fault is a foreseen but unintended consequence of their real purpose. We do not think that such cases should count as manipulation.[24]

As noted, the account we have presented is based on Noggle's, though it is not identical to his. We keep to the spirit of his view that acts of manipulation are "attempts to influence the target's behaviour by getting her to make a mistake," but we depart from the letter, as we outline below.[25] Our account fits well with Noggle's paradigm examples of manipulation: nagging, gaslighting, playing on emotions, exploiting reciprocity, emotional blackmail, and selective attention. These can therefore be used to motivate our view as much as his.

First, nagging induces annoyance in the target, with the aim of making them give in to the nagger's demands in order to stop the constant nuisance of being nagged. It is typically manipulative because the fact that it would stop the annoyance is usually an insufficient reason to justify giving the nagger what they want; the cost of giving in is greater than that of enduring the incessant nagging, so the target intends and acts against their overall reasons. Second, gaslighting is the technique of making someone doubt their own perception and judgment. This is manipulative when there is in fact little or no evidence that there is anything especially wrong with the target's perception or judgment.[26] Third, playing on someone's emotions, as in fearmongering or guilt-tripping, exaggerates certain feelings that the target has little or no reason to feel. When there is no actual danger, inducing fear is manipulative because there is no reason to be fearful and so the fear is a mistake. When the target has in fact done

24 Since we take manipulation to be intentional, we cannot fully account for what makes an act manipulative without giving a general account of intentional action. What relation must hold between the intentional action of inducing an attitude and the belief that said attitude is faulty in order that one intentionally induces a faulty attitude? We defer this question to the scholarship on intentional action, of which there is a substantial body in analytic philosophy. See, for instance, Anscombe, *Intention*; Davidson, "Agency"; Mele, *Springs of Action*; and Setiya, *Reasons Without Rationalism*.

25 Noggle, "Pressure, Trickery, and a Unified Account of Manipulation," 249.

26 Of course, nobody's perception and judgment are perfect, at least not all the time. But a gaslighter exaggerates the imperfections or simply invents them to increase the target's self-doubt.

nothing wrong, inducing feelings of guilt is manipulative because they have no reason to feel guilty. Fourth, exploiting reciprocity is manipulative because it makes the target desire, intend, or feel obliged to do something when in fact they have insufficient reason to do the thing. If I do you a small favor and then ask you to do something for me, depending on you to feel like you owe me, then I am manipulating you, especially if what I ask of you is particularly difficult or costly. Fifth, emotional blackmail, such as the threat of ending a friendship, is manipulative because it makes the target overvalue the friendship—that is, value it more than they have reason to—and so mistakenly decide that complying with the emotional blackmailer's wishes is a comparatively small price to pay. Finally, selective attention makes the target consider only the positive aspects of one possibility and/or only the negative aspects of the other. If successful, this leads the target to pick what they have less reason to choose since they have been paying attention to only the positives of this option and/or the negatives of the other and have therefore overlooked or insufficiently considered some important factors.

The main way in which our account differs from Noggle's is that Noggle splits his account into two types of manipulation based on two types of mistake that might be induced. Manipulation by *pressure* works by offering the target incentives for compliance (positive pressure) or imposing costs for noncompliance (negative pressure).[27] This is distinguished from ordinary (dis) incentivizing by the fact that the pressure is, to use Noggle's term, *nonoptimalizing*—that is, it is not strong enough to make compliance rational. Someone who is manipulated by pressure thus makes the mistake of choosing the lesser good, often due to biases. For instance, someone who succumbs to nagging chooses the lesser good of cessation of annoyance over the greater good of what they previously wanted to (not) do; the nagger has exploited the common human bias towards the short term.[28] Manipulation by *trickery* is getting the target to adopt a faulty mental state—a belief, desire, emotion, etc. that is inappropriate to the situation.[29] Note that our account aligns more closely with manipulation by trickery than pressure, being concerned with faulty mental attitudes. This is because we take manipulation by pressure to be adequately accounted for by manipulation by trickery. Manipulative pressure induces a preference, desire, or intention to choose the lesser good, all of which qualify as faulty attitudes on our description thereof (overall opposed by one's reasons).

27 Noggle, "Pressure, Trickery, and a Unified Account of Manipulation," 244–46, and *Manipulation*, 56–58.

28 Noggle, "Pressure, Trickery, and a Unified Account of Manipulation," 248–49.

29 Noggle, "Pressure, Trickery, and a Unified Account of Manipulation," 243–44, and *Manipulation*, 51–56.

Our version involves just one type of manipulation rather than two, which nonetheless incorporates the same cases as Noggle's. We therefore take our account of manipulation to be both simpler and more precise than Noggle's: simpler because it dispenses with the pressure/trickery distinction and more precise because it includes an account of making a mistake.

One potential objection to our Noggle-inspired account of manipulation might appeal to the view that it is possible to paternalistically manipulate. Our account, it may seem, does not allow for such cases. When paternalistically manipulating, one induces an attitude that the target in fact has most reason to have—at least, that is what the manipulator believes. For instance, it is possible to manipulate someone into having a skeptical attitude toward outlandish claims that are made online concerning vaccines and autism, or the shape of the Earth, or the role of human activity in climate change. Or again, it is possible to manipulate someone into intending to eat healthier foods rather than those that they enjoy more.[30] The thought here is that manipulation is not about the *outcome* of influence but about some other problematic feature and so should not be restricted to cases in which the outcome is bad, wrong, or untrue.[31]

Some influential accounts of manipulation in the current literature focus on the means used by the manipulator rather than on the kind of attitude (or behavior) that they seek to bring about. For instance, there is a class of accounts that focuses on the idea of bypassing rationality.[32] On such a view, manipulation is getting the target to adopt an attitude or behavior on some nonrational basis, bypassing their faculties of rational deliberation. This might involve introducing an additional nonrational psychological influence, such as placing the desired option at the top of a list. Or it might introduce a covert influence that the target does not notice and so cannot deliberate about at all, for instance, through subliminal messages. According to bypassing rationality views or any account that focuses on the process of manipulation rather than its outcome, it is possible to manipulate someone into having an attitude that they have

30 See Thaler and Sunstein, *Nudge*, 1–3, 157–58, 243–46.

31 Gorin criticizes Noggle's view on these grounds in one of his own discussions of manipulation ("Do Manipulators Always Threaten Rationality?" 57–58n14).

32 For examples of this kind of account, see Faden and Beauchamp, *A History and Theory of Informed Consent*, 261, 349; Mele, *Autonomous Agents*, 166–67; Beauchamp and Childress, *Principles of Biomedical Ethics*, 94–95; Baron, "Manipulativeness," 50; Cave, "What's Wrong with Motive Manipulation?" 133; and Blumenthal-Barby, "Between Reason and Coercion," 348–50. For discussion of similar ideas, see Gorin, "Towards a Theory of Interpersonal Manipulation," 89–97, and "Do Manipulators Always Threaten Rationality?"; and Gibert, "The Wrong of Wrongful Manipulation," 339–47.

overall reason to have (e.g., by putting the best option at the top of the list or subliminally supporting true and well-evidenced propositions).

Another kind of account focuses not on the process of manipulation but on the attitude of the manipulator. For instance, Michael Klenk argues that manipulation is negligent influence, characterized by the manipulator's lack of care about whether they are helping the target assess their situation.[33] Maybe a manipulator shows their target reasons or maybe they do not; maybe the attitude or behavior they induce is genuinely rational for their target or maybe it is not. The point is that the manipulator is pursuing their own ends in influencing the target. Like bypassing rationality views, views focused on the attitude of the manipulator do not rule out paternalistic manipulation. The manipulator's agenda may involve the target adopting an attitude that is in fact good or true. If either of these kinds of views are correct, this apparently means that the mistake account, including our variant, is entirely wrongheaded for it is concerned with inducing faulty attitudes—with the outcome. Indeed, the simple fact that there can be paternalistic manipulation may seem to suffice for showing that an outcome-based view such as ours is incorrect.

However, we do not think that paternalistic manipulation is ruled out by our view (or indeed by Noggle's). Consider the case of a doctor who knows that a certain medical treatment is in the best interests of their patient, who is suffering from a certain disease, but the patient is reluctant to take it. The doctor therefore tries to induce greater fear of the disease in the patient so that they will accept the treatment being offered. On the face of it, this appears to be a case of paternalistic manipulation. But consider the following two versions of this situation:

> *Case 1*: The patient's current fear of the disease is appropriate given the severity of the suffering that it will cause them. They are refusing treatment not because they are insufficiently fearful of the suffering but rather because they have the mistaken belief that the treatment is unlikely to be effective—they think it might help but probably will not. The doctor successfully employs certain methods to induce fear in the patient, after which the patient is excessively fearful and is willing to accept the treatment. They still have their mistaken beliefs about the treatment's likely effectiveness, but these are now overcome, from a motivational point of view, by greater fear of the disease.
>
> *Case 2*: The patient's current fear of the disease is inappropriate to the situation. Given the severity of the suffering it will cause them, they are less afraid than they have reason to be. They are refusing treatment

33 Klenk, "Interpersonal Manipulation."

> because they do not fear the disease enough. The doctor successfully employs certain methods to induce fear in the patient, after which the patient is appropriately fearful and is willing to accept the treatment.

In Case 1, the doctor's strategy is manipulative. The patient has an attitude (of fear), induced by the doctor, that is overall opposed to their reasons. This means that paternalistic manipulation is possible on our account; a faulty attitude is induced in order to get the patient to accept what they have overall reason to accept.

What about Case 2? Here, the patient's fear after the doctor's intervention is not itself a faulty attitude so it may seem nonmanipulative on our account. In fact, whether it is manipulative depends on how the doctor induces fear in the patient. If the doctor goes about it by giving the patient detailed information about what the disease will do to them if it goes untreated, what that experience will be like, and what others who have had that disease and gone untreated have said about it, then their fearmongering is not manipulative for they have not induced any mistakes. The beliefs they induce by providing information are not faulty. If, on the other hand, they were to induce the fear by means of promoting some other attitude that is overall opposed by the patient's reasons, then this would be manipulative. For instance, if they gaslit the patient about their previous fear responses, convincing them that they often fear things too little and are consequently far too incautious, then they would be inducing attitudes overall opposed by the patient's reasons. The self-doubt they induce would be a faulty attitude, and therefore the doctor would be manipulating the patient on our account.

The overall point is that paternalistic manipulation involves *indirect* manipulation. The attitudes ultimately brought about might be nonfaulty, such as an appropriate fear of a disease and an intention to accept treatment for it. But the influencer achieves this by inducing what they believe to be faulty attitudes, such as excessive self-doubt. To put the point more generally, suppose that $A$ wants to induce attitude $\alpha$ in $B$, where $\alpha$ is a nonfaulty attitude—that is, an attitude that $B$ has overall reason to adopt. $A$'s method of doing so involves inducing another attitude, $\beta$, in $B$, where $\beta$ is faulty. By our account, $A$ does not thereby directly manipulate $B$ into adopting $\alpha$ but does *use* manipulation to induce $\alpha$. They fulfill the end of inducing a nonfaulty attitude via the means of inducing a faulty attitude.

These are the intuitively correct results. It seems, then, that our version of the mistake account can handle cases of paternalistic manipulation. Now that we have established the account of manipulation with which we will be working, we turn next to what it is to nonmanipulatively influence another. In theory, this

should be straightforward since it is any kind of influence that does not fit the account given above. However, there are certain points that require elucidation.

## 2. NONMANIPULATIVE INFLUENCE

If manipulation is intentionally inducing another to make a mistake, then anything that is not manipulation either does not intentionally induce a mental attitude at all or induces one that the influencer does not believe to be faulty. Since we are here concerned with influence, we can rule out the first possibility; nonmanipulative influence must intentionally induce some attitude in its target. Therefore, it must induce an attitude that the influencer does not believe is faulty. This suggests the following account:

> *A* nonmanipulatively influences *B* to adopt *α* if and only if (1) *A* intentionally brings it about that *B* adopts *α*, and (2) *A* does so while not believing that *α* is a faulty mental attitude.

We could stop there; this is a straightforward account of nonmanipulative influence that follows directly from our account of manipulation. However, we can make this more precise by clearing up two potential ambiguities.

First, not believing that the attitude is faulty is not the same as believing that the attitude is not faulty. The influencer might instead suspend judgment on this point, either because they do not care to consider it or because they cannot come to a conclusion on the matter. Take the doctor in Case 2 above. Suppose they provide testimonials about how bad the disease can become in order to induce the proper degree of fear in their patient. Suppose further that these testimonials are, in the objective sense, reasons for the patient to be afraid of the disease. The doctor might believe (rightly) that the patient's resulting attitude of fear is nonfaulty, in which case it is quite intuitive that they are not manipulating the patient. However, they might alternatively have no such belief but just think that this is an effective technique for getting the patient to accept treatment. Perhaps they simply think about it not in terms of objective reasons or (non)faulty attitudes but just in terms of effectiveness. Or maybe they have considered whether the fear they are trying to induce would be faulty but are unsure on the matter. Either way, they are not being manipulative on the account we use. Only if the doctor positively believes (mistakenly, in this case) that the patient's fear is faulty are they manipulating the patient.

On a looser definition, we might say that an influencer who suspends judgment about their target's reasons risks accidentally or negligently manipulating their target. Indeed, some accounts in the literature explicitly take this approach. Moti Gorin suggests that presenting what one takes to be objectively

good reasons is manipulative if the influencer does not care about the quality of the reasons, only about getting the desired result.[34] Similarly, Klenk takes manipulation to be careless influence, wherein the manipulator simply does not care whether their method of influence reveals reasons to their target.[35] The account we favor requires that manipulation be intentional—specifically, intentionally inducing a mistake—so it cannot be done accidentally or negligently, though we acknowledge that this may not always be intuitive.

Second, we should clarify what it means for reasons to be overall neutral with regard to an attitude—that is, for one's reasons to neither overall support it nor overall oppose it. It might at first glance be thought that neutrality implies that there are no strong reasons either for or against it, but this is not the case. When there are rationally incompatible possible attitudes, such as believing that $p$ and believing that $\neg p$, or intending to $\phi$ and intending to not $\phi$, any reason favoring one disfavors the others.[36] Therefore, when there is a unique most supported attitude, each of the other potential attitudes are overall opposed by reasons. Only when there are two or more attitudes that are equally most supported by reasons is there neutrality. For instance, if you know that Smith is either in London, Oxford, or Paris, and the only evidence pertaining to which suggests that Smith is in England, then your reasons are overall opposed to believing that Smith is in Paris and are neutral with respect to believing that Smith is in London and believing that Smith is in Oxford.

To state it more precisely, reasons are overall neutral with respect to an attitude just in case there is no rationally incompatible attitude better supported by the reasons and there is at least one rationally incompatible attitude that is equally well supported by the reasons (joint best); an attitude is overall supported by reasons just in case there is no rationally incompatible attitude either better supported or equally well supported by the reasons (unique best). In order to be a faulty attitude—a mistake—an attitude need not be severely disfavored by the reasons. There merely needs to be at least one rationally incompatible attitude that is better supported by one's reasons.

## 3. PERSUASION

### *3.1. An Account of Persuasion*

We have provided an account of nonmanipulative influence based on Noggle's account of manipulation. Our main purpose, however, is to account for

34 Gorin, "Do Manipulators Always Threaten Rationality?" 57–58.

35 Klenk, "(Online) Manipulation," 97.

36 See, for instance, Snedegar, *Contrastive Reasons*, 34–36.

persuasion, which for current purposes, is assumed to be a specific kind of nonmanipulative influence.

Persuasion is a kind of nonmanipulative influence, but it is not the only kind. Another important kind of nonmanipulative influence is incentivization, which we assume can involve the provision of either positive incentives or negative incentives (i.e., disincentives) and which includes the category of coercion.[37] Recall the discussion of manipulative pressure, which we classified as a type of manipulative trickery. Manipulative pressure provides *nonoptimalizing* incentives and disincentives; it is irrational to give in to such pressure, but due to standard human biases and weaknesses, people often do. However, it is possible for the pressure to be significant enough that changing one's mind is the rational thing to do.

If the pressure of an incentive is *optimalizing*—that is, it is a sufficiently great positive or negative incentive that it is rational for the target to accept—then it is nonmanipulative.[38] Here, we consider optimalizing—and therefore nonmanipulative—incentives. They can be morally dubious, as coercion generally is, but they need not be. Examples may include rewards, prizes, punishments, reciprocation, and so on. The incentivizer adjusts the target's circumstances in order to get them to adopt an attitude, rather than trying to change their mind directly.[39] This is what distinguishes incentivization from persuasion. Persuading someone to choose one option over another is precisely a matter of changing their mind rather than of changing their circumstances. One might, as part of the persuasion, provide them with information, appeal to their better (or worse) nature, try to show that some considerations are more important than others, or point out the likely consequences of their actions, but actually altering the options or the situation in which the choice is made is beyond the boundaries of persuasion.

The difference is akin to the important distinction between issuing a threat and issuing a warning, even a warning about one's own future conduct.[40] Both warnings and threats are typically conditional claims made in order to alter someone's behavior by making them believe that their current (intended)

37 We do not take a stand on exactly which cases of incentivization count as coercion.

38 Noggle, "Pressure, Trickery, and a Unified Account of Manipulation," 248.

39 Noggle points out that it can be rational to resist coercion for it sends the message that one will not give in to such methods, which, in the long term, might be beneficial. However, coercion always makes compliance rational when the current choice is taken in isolation. See Noggle, "Pressure, Trickery, and a Unified Account of Manipulation," 247. Presumably, similar points apply to other optimalizing incentives.

40 See, for instance, Nozick, "Coercion"; Greenawalt, "Criminal Coercion and Freedom of Speech"; and Sachs, "Why Coercion Is Wrong When It's Wrong."

behavior will have bad consequences for them. The key difference is that a warning is concerned with the situation *as it stands*, whereas a threat is concerned with how the situation may be *changed* if they continue. For instance, "If you lean on that barrier, it will give way and you will fall" is a warning, whereas "If you lean on that barrier, I will push you over it" is a threat. An intuitive way of marking the distinction is to say that with a threat, the consequence is an act to be undertaken by the threatener (or perhaps an agent working on their behalf). But that is not all. For it is possible to issue a warning about one's own action that does not amount to a threat. For instance, if you have a conditional intention to do something in certain circumstances, or certain triggers will make you act on powerful compulsions, or you have a duty that you strictly adhere to, then you might issue a warning—not a threat—about your intention, compulsion, or duty. You might warn a wayward partner that you will leave them if they are unfaithful; you might warn an aggressive stranger that you will be compelled to defend yourself if they come near you; you might tell a friend that you will feel duty bound to report them to the authorities if they proceed with some illicit confidence trick. In each of these cases, although your own actions are the consequences being warned of, you are not *threatening* them in the sense mentioned above. The conditional proposition of, say, "If you proceed with the scam, I will tell the police" was already true but is only now becoming salient. You are not changing the situation to get them to change their behavior but merely informing them of the consequences should they act in a certain way.

In practice, it can be very difficult to tell the difference between warning and threatening, just as it can be difficult in practice to tell the difference between persuasion and incentivization since it depends in part on the intentions of the one issuing the threat or warning. One test to tell them apart is this: if you would follow through with the consequences (e.g., report them to the police) even if you had not told them that you would do so, then it is a warning; if you would follow through only if you had told them, then it is a threat.[41] The warning does not change the situation—the conditional claim about your actions is true whether you tell them or not—whereas the threat does change the situation, making it the case that the conditional claim is true. In our estimation, warning is a type of persuading, as we will demonstrate once the full account of persuasion is laid out. Successfully threatening someone is typically a kind of incentivization (and often coercion) as it makes it rational for them to change their intentions. However, it can also be a kind of manipulation, either by being an empty threat (and so be a case of trickery) or by being insufficiently strong

41 Greenawalt, "Criminal Coercion and Freedom of Speech," 1097–98.

to make the change rational and relying on the target to make the mistake of complying anyway (and so be a case of manipulative pressure).

The idea that persuasion changes someone's mind and incentivization changes their situation is intuitive, but to properly make the distinction, we need to be more precise. All of the above examples of what persuasion can involve—providing information, appealing to the target's better nature, showing the relative importance of different considerations, and so on—change the circumstances of the target to some extent. Their circumstances now include having more information than before or being more conscious of a conflict with their own values or understanding better the weight to be given to various considerations. Indeed, any instance of presenting someone with a reason changes their circumstances. So the distinction between incentivization and persuasion cannot straightforwardly be made on the basis of changing circumstances versus changing minds.

For both—persuasion and incentivization—the influencer tries to induce an attitude in their target that they do not believe is overall opposed to the target's reasons, so both qualify as nonmanipulative influence. The crucial difference is that when incentivizing, *A* tries to make *a* more well supported by *B*'s reasons by offering a reward or threatening a penalty—that is, by adding positive or negative pressure. When persuading, *A* does not make *a* any more well supported by *B*'s reasons but rather draws *B*'s attention to the preexisting reasons that support *a*. We suggest that given the conditions that would make the attitude in question not faulty (i.e., there is no attitude with which it is rationally incompatible better supported by reasons), persuasion *shows* the content of the attitude to meet those conditions, as opposed to *making* the content of the attitude meet those conditions. So persuasion, intuitively understood as changing someone's mind but not their circumstances, can be more precisely understood as *showing* that the relevant attitude is not faulty—either supported by or neutral with respect to the target's reasons. Incentivization, intuitively understood as changing someone's circumstances to such a degree that the given attitude becomes rational, can be more precisely understood as *making it the case* that the attitude is not faulty. For instance, the incentivizer makes it the case that an action is a means to the end of gaining a reward or avoiding a penalty.[42]

42 One upshot of this is the fact that the scope of influence that targets circumstances rather than minds is broader than just coercion and incentivization, which are principally focused on conative attitudes like desire and intention. Making some proposition true and well evidenced makes a belief with that proposition as its content nonfaulty, but doing so is not incentivizing or coercing someone into having such a belief. This may not be a typical method of influence but is clearly not a kind of persuasion and is worth distinguishing from it.

There is another crucial distinction to be made concerning reasons for an attitude: that between the reasons that *there are* and the reasons that *one has*.[43] For instance, an agent may not know all the facts that are reasons. Alternatively, they may know some fact without recognizing it as a reason for adopting a certain attitude. Or again, they may know a given fact and recognize it as a reason but not assign the appropriate weight to it. Each of these is a way in which there can be a reason without the given agent having it, since they do not have appropriate epistemic access to the reason. In such cases, correcting them, whether by informing them of the relevant facts, showing that a fact is a reason for an attitude, or showing the appropriate weight to assign to a reason, is not manipulative but rather persuasive. One way of distinguishing persuasion from other nonmanipulative influences is therefore to say that persuasion provides its target with appropriate epistemic access—access that facilitates the transition from *there being* reasons to the target *having* those reasons. Henceforth, unless otherwise stated, when we talk of someone's reasons for an attitude, we mean the reasons that *there are* for them to have that attitude rather than the reasons that *they have*.

So persuasion is nonmanipulative influence that induces an attitude by means of reason giving, where reason giving is understood as showing that the overall weight of the target's reasons supports the given attitude to a greater extent than the target previously realized. Furthermore, the persuader induces the attitude in their target without the positive belief that it is a mistake for said target. But that does not mean that persuasion, even understood in this nonmanipulative sense, is always innocuous. For example, there are cases of persuasion in which the persuader suspends belief on whether the resulting attitude is faulty, but it is negligent of the persuader to do so since there is in fact sufficient evidence available that the attitude *is* faulty. Or again, the persuader may have an ulterior motive for persuading their target. If it will be in some way beneficial to an influencer for their target to adopt a certain attitude, and it just happens that this would be a nonfaulty attitude and presenting reasons is the most efficient way of making them adopt it, this is a case of persuasion in our view. In some cases, persuasion is therefore morally problematic and indeed even impermissible.[44] We consider this to be an advantage for there is intuitively such a thing as impermissible persuasion. But it also appears to go against one of the main purposes of distinguishing persuasion from manipulation in the first place: manipulation is presumptively impermissible, whereas

43 Kaiserman, "Reasons-Sensitivity and Degrees of Free Will," 689.

44 Indeed, the examples just given would be classed as manipulation by, respectively, Klenk, "Interpersonal Manipulation" and "(Online) Manipulation," 96–97; and Gorin, "Do Manipulators Always Threaten Rationality?" 56–59.

persuasion is presumptively permissible. If some cases of persuasion are impermissible, then perhaps we have drawn the boundary in the wrong place.

However, even impermissible persuasion is to be distinguished from manipulation. When persuasion is impermissible, it is not so in virtue of being persuasive, whereas impermissible manipulation is often impermissible in virtue of being manipulative. Our account of persuasion does not rule out impure motives on the part of the persuader or negative consequences for the target, either of which may render acts of persuasion impermissible. Manipulation, on the other hand, does not require impure motives (beyond the intention to induce a faulty attitude), negative consequences, or any other factor outside of what makes the act manipulative to render it immoral. On the contrary, it stands in need of moral justification if it is to be permissible. For instance, as we have seen, it can be used as part of a paternalistic strategy, which may sometimes make it, all things considered, the right thing to do. All this is to say that persuasion is *presumptively* permissible, whereas manipulation is *presumptively* impermissible. Here, then, is our account of persuasion:

> *A* persuades *B* to adopt *α* if and only if (1) *A* intentionally brings it about that *B* adopts *α*, (2) *A* does so only by presenting (some of) what they believe to be the reasons that there are for *B* to adopt *α* (as opposed to the reasons that *B* has for adopting *α*), and (3) *A* does so while not believing that *α* is a faulty mental attitude.

This account has several important payoffs. First, it analyzes an important concept that has so far received little direct philosophical attention. Much more attention has been given to other kinds of morally significant influence, such as manipulation and coercion; persuasion has often remained in the background. Second, it fits well with an influential account of manipulation (that of Noggle) and, more generally, fits into a broader taxonomy of influence based on the faultiness or otherwise of attitudes and their inducement by changing circumstances or by changing minds. Third, just as Noggle's account of manipulation provides a ready explanation of why manipulation is presumptively wrong (it induces an attitude that the influencer takes to be mistaken), our account of persuasion provides a ready explanation of why persuasion is presumptively permissible (it induces an attitude that the influencer does not take to be mistaken) while at the same time allowing that persuasion can sometimes be wrong even without being manipulative. Finally, the distinction between persuasion and incentivization maps onto and generalizes from another important conceptual distinction: that between warning and threatening. Both are distinctions based on the difference between making something the case and showing that it is the case.

*3.2. Dowding's Account*

A further virtue of our account is that it compares favorably with one of the main well-developed accounts of persuasion currently in the philosophical literature. Dowding takes persuasion to be a speech act that leads someone to change their beliefs about something.[45] He then argues that nonmanipulative noncoercive persuasion is that that meets two "reliability conditions": first, the speaker's reasons for believing what they assert must be the same as the hearer's reasons for believing it; second, the speaker's intention must be that they and the hearer learn the truth together. Our account, assuming, as we do, that persuasion is mutually exclusive of both manipulation and incentivization (including coercion), is more readily comparable with persuasion that meets these two conditions than with what Dowding calls persuasion, which is what we would call *influence on belief*.

This exclusive focus on belief is the first problem with Dowding's account. There are many other attitudes that one can be persuaded to have, such as persuading a patient to fear a disease to a greater extent or persuading a friend to intend to stay at home rather than go on holiday. Could Dowding's account be adapted to capture other attitudes as well? Not easily, we think. The reliability conditions require that the persuader themself has the belief that they are trying to inculcate in their target and that their intention is for them to learn the truth together. A modified version of Dowding's account would therefore struggle to account for cases in which the resulting attitude is supported by reasons that are specific to the persuadee. Take Case 2 presented earlier, the case of a doctor inducing an appropriate degree of fear in their patient of their disease. By our account, provided the doctor restricts their methods of influence to presenting reasons, they are persuading the patient. However objective and well informed the doctor is, they will not fear the patient's disease either as much or in the same way as the patient ought to. The reasons for fearing the disease—that it will cause suffering, that it will necessitate unpleasant medical procedures, that it will disrupt their lifestyle, and so on—apply to the patient specifically. The doctor would not themself have the attitude that they are trying to inculcate, nor is it their intention that they and the patient attain "true/appropriate fear" together. Similar remarks can be made of persuading someone to intend or desire something; reasons for such attitudes, while they may be objective in the sense specified above—that is, they are facts—sometimes apply only to certain people. Our account, which is not limited to persuasion of beliefs, does not have these problems. It allows for reasons to be subject relative when such reasons are appropriate to the attitude.

45 Dowding, "Power and Persuasion."

There are further problems with each of the two reliability conditions. Dowding's first reliability condition—that the speaker's reasons for believing what they assert must be the same as the hearer's reasons for believing it—is comparable to our condition 2—that the persuader presents only what they believe to be their target's reasons. However, Dowding is too restrictive in requiring the reasons of the hearer to be exactly those of the speaker. If a speaker gives only some of their reasons and thereby convinces a hearer, is that not a case of persuasion? When there is limited time, one might leave out what one takes to be weaker reasons, presenting only the most persuasive. One does not thereby fail to perform an act of persuasion.

The second reliability condition—that the speaker's intention must be that they and the hearer learn the truth together—is comparable to a combination of our conditions 1 and 3—that the persuader intentionally induces the target to adopt the attitude and that the persuader does not believe that attitude to be faulty. However, it is quite standard (though, as argued above, not essential) for a persuader to take themselves to have already ascertained the truth; their intention in such cases is for the hearer to learn the truth that they themselves have already learned. Therefore, they do not need to intend their persuasive act as part of a joint search for truth. What Dowding seems to miss is that persuasion can be—and typically is—one-way.

Dowding's account is most readily interpreted as an account of morally ideal persuasion of a belief. It requires the persuader to be acting in good faith (holding the belief in question themself) and from pure motives (learning the truth together with their target). However, we think it important to take a broader approach. As shown, our account encompasses the many and varied attitudes that someone can be persuaded into adopting, as well as the more morally dubious (though not manipulative) kinds of persuasion.

### *3.3. Applying the Account*

We are now in a position to reconsider the unclear examples that were mentioned at the start of this article and assess them in the light of our accounts of persuasion and manipulation. Whether they count as persuasion or manipulation depends on the details of each case, and we can now determine which details are relevant. The key factors that distinguish persuasion from manipulation are that (1) the persuader's method is restricted to presenting reasons and (2) the persuader does not believe that the attitude in question is faulty. These, then, are the points to focus on.

The first example involves telling a friend that the friendship will be over if they do not change their ways. For this to count as persuasion rather than manipulation, it first needs to be the case that you really do have the conditional

intention of withdrawing friendship and that this intention has not been formed simply for the purposes of changing your friend's mind. That is to say, that you will withdraw friendship is a reason—a fact that bears on the matter—and it is a reason being presented rather than created. Second, it must be the case that you do not believe that your friend changing their ways is overall opposed by your friend's reasons. That is to say, you must either believe that, taking all the friend's reasons into account, changing their ways (and maintaining the friendship) is objectively at least as good an option as maintaining their current ways (and losing the friendship), or suspend judgment on the matter. If these conditions are fulfilled, then it is a case of persuasion. If the second is not fulfilled, then it is manipulation. If the first is not fulfilled but the second is—for instance, you use the withdrawal of a hitherto valuable friendship as a bargaining chip rather than as a principled act of "drawing the line"—then it is incentivization.

The second example involves persistently telling your friend to change their ways and why you think they should do so, which might be construed as nagging. Again, we can say that this counts as persuasion if you believe that complying is not opposed by their reasons, and you restrict your method to presenting reasons.

The third example involves appealing to your friend's moral emotions and sense of decency, which may be construed as emotional manipulation. This also depends on the details. So long as you believe that the weight of reasons—including, in this case, moral reasons—does not disfavor your friend making the change you suggest, and your only method is to present those reasons, then it is persuasion. However, if you, for instance, exaggerate the importance of the moral reasons, then you are inducing them to make a mistake, so that is a case of manipulation.

Finally, we raised the matter of very minor influences, such as dressing smartly, smiling, and discussing an interlocutor's favorite topics in order to make a good first impression. Are these manipulative? Again, the answer varies depending on the details. These behaviors may sometimes give reasons. For instance, dressing smartly gives someone reason to think you are organized and understand social niceties, which may in turn give them reason to think you competent in various respects. Discussing an interlocutor's favorite topic gives them reason to consider you a versatile conversationalist who is willing to adapt. It may also give them reason to like you and to want to spend time with you. But in other cases, these techniques may be used without either the intention or the effect of presenting reasons. But even when not persuasive, they might not be manipulative. If you believe that whatever attitudes constitute a good first impression—enjoyment of your company, belief that you are a decent and/or competent person, desire to see more of you, etc.—really

are supported by their reasons, then you are not inducing them to make what you believe to be a mistake by so doing, even if you are not presenting reasons. What this kind of influence is properly called is an interesting question, but not one on which we will take a stand here. On the other hand, if you believe that the attitudes that constitute a good first impression are opposed by their reasons, then even these minor influences count as manipulation. For instance, using them to convince someone that you really do have their best interests at heart when in fact you do not is manipulative. Even small and seemingly trivial acts can therefore carry moral weight if performed in certain circumstances.

## 4. CONCLUSION

We have put forward a relatively simple and intuitive account of persuasion based on the assumption that persuasion is not manipulative. Drawing on Noggle's central idea that manipulation is inducing a mistake, we have argued that persuasion is inducing, only by presenting reasons, an attitude not believed by the influencer to be mistaken. As mentioned at the beginning, we are open to the possibility that persuasion in fact overlaps with manipulation. If that is the case, then our account is of specifically nonmanipulative persuasion; some broader account would then be needed to encapsulate persuasion as a whole. However, even if this is the case and our account therefore captures only a strict subset of cases of persuasion, we think that the subset it captures is nonetheless philosophically interesting and worth delineating. Manipulation and persuasion are important categories of methods of influence, so identifying a boundary between them is a significant step in gaining a greater understanding of the ways we can influence one another. Furthermore, as we have highlighted, persuasion is a concept that has had relatively little philosophical attention so the value of our proposed account lies also in advancing an underdeveloped topic.

Having set out and argued for the account, we have then demonstrated a number of important payoffs that our account has, including its advantages over an account of persuasion already in the literature. Notably, our account can explain why persuasion is presumptively permissible, whereas manipulation is not, while still allowing for morally dubious kinds of persuasion. It can also make sense of several "borderline" cases in which it is unclear whether the influencer is persuading or manipulating their target.

It seems to us that there are resources here to expand the project further, to give a fuller taxonomy of many kinds of influence in terms of whether they change minds or circumstances and whether they make an attitude rational or not. Such a taxonomy would not only help make sense of manipulation, which refers to such a broad variety of techniques that some suggest it is a

"family resemblance" concept rather than one specific phenomenon.[46] It would also be fruitful for distinguishing among different kinds of persuasion and incentivization.[47]

*Institute for Ethics in AI*
*Corpus Christi College, University of Oxford*
*thomas.mitchell@philosophy.ox.ac.uk*

*Uehiro Oxford Institute*
*Jesus College, University of Oxford*
*thomas.douglas@uehiro.ox.ac.uk*

## REFERENCES

Anscombe, G. E. M. *Intention*. Basil Blackwell, 1957.

Baron, Marcia. "Manipulativeness." *Proceedings and Addresses of the American Philosophical Association* 77, no. 2 (2003): 37–54.

Beauchamp, Tom, and James Childress. *Principles of Biomedical Ethics*. 5th ed. Oxford University Press, 2001.

Blumenthal-Barby, J. S. "Between Reason and Coercion: Ethically Permissible Influence in Health Care and Health Policy Contexts." *Kennedy Institute of Ethics Journal* 22, no. 4 (2012): 345–66.

———. "A Framework for Assessing the Moral Status of 'Manipulation.'" In Coons and Weber, *Manipulation*.

Brown, Rebecca. "'Sneaky' Persuasion in Public Health Risk Communication." *Ratio* 38, no. 4 (2024): 195–278.

Buss, Sarah. "Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints." *Ethics* 115, no. 2 (2005): 195–235.

46 For examples of the family resemblance view, see Todd, "Manipulation," 4; Coons and Weber, "Manipulation," 6–8; Blumenthal-Barby, "A Framework for Assessing the Moral Status of 'Manipulation'"; Gibert, "The Wrong of Wrongful Manipulation," 365–66; and Cohen, *The Concept and Ethics of Manipulation*, 63–71.

47 The work for this article was supported by a Consolidator Award from the European Research Council (grant number 819757) and by funding from the Uehiro Foundation on Ethics and Education. We would like to thank two anonymous reviewers from the *Journal of Ethics and Social Philosophy* for insightful criticism that has enabled us to substantially improve on earlier versions of this article. We would also like to thank participants at the 2024 Joint Session of the Aristotelian Society and the Mind Association for hearing and commenting on the ideas from which this article grew.

Cave, Eric. "What's Wrong with Motive Manipulation?" *Ethical Theory and Moral Practice* 10, no. 2 (2007): 129–44.

Cohen, Shlomo. *The Concept and Ethics of Manipulation*. Cambridge University Press, 2025.

———. "Manipulation and Deception." *Australasian Journal of Philosophy* 96, no. 3 (2018): 483–97.

Coons, Christian, and Michael Weber. "Manipulation: Investigating the Core Concept and Its Moral Status." In Coons and Weber, *Manipulation*.

———, eds. *Manipulation: Theory and Practice*. Oxford University Press, 2014.

Crisp, Roger. "Persuasive Advertising, Autonomy, and the Creation of Desire." *Journal of Business Ethics* 6, no. 5 (1987): 413–18.

Davidson, Donald. "Agency." In *Essays on Actions and Events*. Oxford University Press, 1980.

Douglas, Thomas. *Protecting Minds: The Right Against Mental Interference*. Oxford University Press, 2026.

Dowding, Keith. "Power and Persuasion." *Political Studies* 64, no. 15 (2016): 4–18.

Faden, Ruth, and Thomas Beauchamp. *A History and Theory of Informed Consent*. Oxford University Press, 1986.

Fischer, John, and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press, 1998.

Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66, no. 23 (1969): 829–39.

Frege, Gottlob. "Sense and Reference." *Philosophical Review* 57, no. 3 (1948): 209–30.

Gibert, Sophie. "The Wrong of Wrongful Manipulation." *Philosophy and Public Affairs* 51, no. 4 (2023): 333–72.

Gorin, Moti. "Do Manipulators Always Threaten Rationality?" *American Philosophical Quarterly* 51, no. 1 (2014): 51–61.

———. "Towards a Theory of Interpersonal Manipulation." In Coons and Weber, *Manipulation*.

Greenawalt, Kent. "Criminal Coercion and Freedom of Speech." *Northwestern University Law Review* 78, no. 5 (1983): 1081–124.

Kaiserman, Alex. "Reasons-Sensitivity and Degrees of Free Will." *Philosophy and Phenomenological Research* 103, no. 3 (2021): 687–709.

Klenk, Michael. "Interpersonal Manipulation." Unpublished manuscript, June 2021. https://ssrn.com/abstract=3859178.

———. "(Online) Manipulation: Sometimes Hidden, Always Careless." *Review of Social Economy* 80, no. 1 (2022): 85–105.

Mele, Alfred. *Autonomous Agents: From Self-Control to Autonomy*. Oxford Uni-

versity Press, 1995.

———. *Springs of Action: Understanding Intentional Behavior*. Oxford University Press, 1992

Mercier, Hugo, and Dan Sperber. "Why Do Humans Reason? Arguments for an Argumentative Theory." *Behavioural and Brain Sciences* 34, no. 2 (2011): 57–74.

Mitchell, Thomas, and Thomas Douglas. "Wrongful Rational Persuasion Online." *Philosophy and Technology* 37, no. 1 (2024).

Noggle, Robert. *Manipulation: Its Nature, Mechanisms, and Moral Status*. Oxford University Press, 2025.

———. "Pressure, Trickery, and a Unified Account of Manipulation." *American Philosophical Quarterly* 57, no. 3 (2020): 241–52.

Nozick, Robert. "Coercion." In *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, edited by Sidney Morgenbesser, Patrick Suppes, and Morton White. St. Martin's Press, 1969.

Pereboom, Derk. *Living Without Free Will*. Cambridge University Press, 2001.

Raz, Joseph. *Practical Reason and Norms*. 3rd ed. Hutchinson, 1999.

Sachs, Benjamin. "Why Coercion Is Wrong When It's Wrong." *Australasian Journal of Philosophy* 91, no. 1 (2013): 63–82.

Scanlon, T. M. *What We Owe to Each Other*. Belknap Press, 1998.

Setiya, Kieran. *Reasons Without Rationalism*. Princeton University Press, 2009.

Snedegar, Justin. *Contrastive Reasons*. Oxford University Press, 2017.

Thaler, Richard, and Cass Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, 2008.

Todd, Patrick. "Manipulation." In *International Encyclopedia of Ethics*, edited by Hugh LaFollette. John Wiley and Sons, 2013.

Tsai, George. "Rational Persuasion as Paternalism." *Philosophy and Public Affairs* 42, no. 1 (2014): 78–112.

Williams, Bernard. "Internal and External Reasons." In *Moral Luck: Philosophical Papers, 1973–1980*. Cambridge University Press, 1981.