Ethical Theory and Moral Practice
https://doi.org/10.1007/s10677-024-10458-9

Influencer-Centered Accounts of Manipulation

Micha H. Werner[1]

Accepted: 26 July 2024

Abstract
Advances in science and technology have added to our insights into the vulnerabilities of human agency as well as to the methods of exploiting them. This has raised the stakes for efforts to clarify the concept and ethics of manipulation. Among these efforts, Robert Noggle's influencer-centered account of manipulation has been most significant. He defines manipulative acts as those whereby an agent intentionally influences a recipient's attitudes so that they do *not* conform as closely as they otherwise would to the pertinent norms and ideals endorsed by the *influencer*. This provides a relatively simple and in many ways clear definition of manipulation. It sidesteps thorny debates about autonomy, freedom, or practical rationality. It also promises to reveal a conceptual parallel between manipulating and lying, and thus to explain why manipulation is *pro tanto* wrong. In one respect, however, the account remains ambiguous: It remains unclear whether, and to what extent, it requires that influencers' beliefs about what is ideal *for their recipients* to be grounded in some effort on the part of the influencer to identify with or take on the role of her recipient. This paper explains this ambiguity. It argues that influencer-centrism cannot remain indifferent to the *validity* of an agent's beliefs about the ideal state of the recipient *and* provide an identification requirement that would render the whole account plausible and sufficiently determinate.

Keywords Autonomy · Lying · Manipulation · Robert Noggle · Role-taking

Recent events such as the proliferation of COVID conspiracies have put a spotlight on issues of manipulation. These issues arise from a broader undercurrent of scientific, technological and social trends: Digitization has made it easier to "flood" the information space, sow doubt, and impede access to appropriate content (Bennett and Livingston 2020). Micro-targeted mass communication allows information to be tailored to individual vulnerabilities (Zuiderveen Borgesius et al. 2018; cf. Jongepier and Wieland 2022). Social networks enable

✉ Micha H. Werner
micha.h.werner@gmail.com

[1] Department of Philosophy, University of Greifswald, 17489 Greifswald, Germany

Published online: 13 August 2024

'influencers' and their principals to harness the forces of group behavior even when audiences are dispersed over large distances (Sherman et al. 2018; Vrontis et al. 2021; Hudders and Lou 2022). At the same time, studies in behavioral economics, psychology and neuroscience provide potential manipulators with ever more specific insights into our epistemic and motivational vulnerabilities; and while the same insights might also be useful in coping with these vulnerabilities, some proposed coping mechanisms ('nudges') are themselves suspected of being manipulative (cf. Thaler and Sunstein 2021; Hanna 2015; Noggle 2018; Sunstein 2016).

In line with the increased urgency of the issue, philosophical work on the concept and ethics of manipulation has intensified (Noggle 1996, 2018, 2020a, b; Baron 2003; Buss 2005; Cave 2007; Coons and Weber 2014b; Fischer 2017; Fischer and Illies 2018; Susser et al. 2019a, b; Klenk 2020, 2022; Jongepier and Klenk 2022; Turza 2023). Among the more recent literature, Robert Noggle's contributions are the most consequential. In his 1996 paper, he proposes to understand manipulative acts as intentional efforts to 'lead astray' (Noggle 1996): Manipulators try to prevent their recipient's beliefs, desires, or emotions from conforming (more closely than they would otherwise) to the relevant norms or ideals endorsed by the manipulators. This account has profoundly shaped the subsequent discussion, and has been defended and further developed by various authors.[1] Its impact is hardly surprising: Noggle starts his discussion from a list of well-chosen examples that represent quite different forms of influence. The account sidesteps thorny questions about autonomy or practical rationality. It promises to avoid controversial commitments regarding the content and validity of the norms and ideals on which the manipulativeness of actions depends. It aims to uncover a similarity between manipulating and lying, and thereby to explain why manipulative acts are *pro tanto* wrong. The account also seems comparatively simple: Noggle himself notes that it explains the moral significance of manipulation in a "fairly simple" way (Noggle 1996, p. 52).

A core element of the account remains ambiguous, however. This paper explains this ambiguity and discusses its implications. It begins with an outline of the influencer-centered account (1) and then focuses on one of its core notions: the influencer's beliefs about what is good for her recipient (2). Drawing on examples and arguments from Noggle, Barnhill (2014), Manne (2014), and Hanna (2015), it then argues that the account has counter-intuitive implications which result from its indifference about the validity of those beliefs (3). After considering possible defenses, it argues that the same indifference limits the account's ability to provide first-personal guidance to influencers who are concerned about manipulating others (4). It concludes that the apparent simplicity of influencer-centrism results in part from its being under-determined, and that its basic structure may even complicate efforts to develop a more determinate account (5).

[1] For example, Barnhill declares that "[t]he core of Noggle's account of manipulation is correct" (Barnhill 2014, p. 65) and provides insightful modifications (see also Barnhill 2016, p. 314, 2022), Gorin states that his account of manipulative action "is similar to Robert Noggle's" (Gorin 2014a, p. 95 n. 28) as well as to Barnhill's (Gorin 2022, p. 202). The influence of Noggle's account is also visible in important contributions that aim to provide an alternative; for a broader assessment of its impact see Coons and Weber (2014a, p. 11Noggle (2020a).

1 Noggle's Account of Manipulation

Noggle's conceptual exploration starts from a list of eight examples of interpersonal influence. These include instances of lying to, deceiving, and misleading others, as well as instances of overtly tempting recipients (Noggle 1996, p. 43) or playing on their emotional weaknesses or lack of self-control. Noggle states that "all" eight cases "seem like cases of manipulation"[2] and develops an account of manipulation that is meant to cover all of them. According to this account, manipulative acts are intentional efforts of an agent to lead the recipient "astray" (Noggle 1996, p. 44) by exerting a specific kind of influence on her beliefs, desires, or emotions. In a fittingly mechanistic metaphor, Noggle refers to the recipient's beliefs, desires, and emotions as "levers" that a potential manipulator can "operate" (Noggle 1996, p. 44).

Noggle's criterion for manipulative influence then draws on the idea that "there are certain norms or ideals that govern beliefs, desires, and emotion": *Beliefs* should be true and, moreover, "relevant to the situation at hand" (Noggle 1996, p. 44). *Desires* should be rational in that they "conform to our beliefs about what we have reason to do", which implies that they should not be distorted by processes of psychological conditioning or canceled out by weakness of the will (Noggle 1996, p. 45). Even *emotions* are subject to certain norms and ideals: Since they serve to make "salient whatever is most important, most relevant to the situation at hand", they can be more or less "appropriate" (Noggle 1996, p. 45 f.).

Whether intentionally influencing a recipient's beliefs, desires, or emotions is manipulative now depends on the *direction* of the influence *relative to the norms and ideals* to which those beliefs, desires, or emotions are susceptible. An action is manipulative if and only if an agent thereby intentionally influences a recipient's attitudes so that they do *not* conform as closely to the relevant norms and ideals as they otherwise would. To use Noggle's metaphor, a manipulator pulls (or perhaps sometimes just 'holds steady') the "levers" of the recipient's beliefs, desires, or emotions "away" from the ideal "lever setting". Because of the crucial role of norms and ideals in identifying manipulation, Gorin (2022, p. 199) calls Noggle's account "norm-based".

This is where things get more complicated. For Noggle recognizes that in many cases, it may be debatable which settings of "someone's internal levers" could be considered "ideal". For example, the "notions of relevance and appropriateness" (Noggle 1996, p. 47) that Noggle considers important for evaluating beliefs and emotions are often contested.

Just to be clear, we may note that there are actually *two* levels of potential controversy. On a fundamental level, one might challenge Noggle's own reconstruction of the *general* norms and ideals for beliefs, desires, and emotions. For example, one might question the idea that the adequacy of emotions should (mainly) be measured by their ability to make "salient whatever is most important […] to the situation at hand". On a more *specific* level, there may be disagreement about *how to apply* the general norms and ideals to the concrete situation of a recipient at the time where she is affected by an attempt to influence her.

[2] This remains controversial especially with regard to cases of *overt* influence. While Barnhill (2014, pp. 59–60), Gorin (2014a, pp. 74–81, b, pp. 80–81), and Klenk (2022) share Noggle's (1996, p. 43) conceptual intuition, Susser, Rössler and Nissenbaum (2019a, p. 19) argue that instances of overt influence should rather be considered as cases of either persuasion (albeit perhaps by bad reasons) or coercion. Baron (2003, p. 44) seems to take a middle ground, arguing that minute amounts of dishonesty may suffice to render expressions of emotion manipulative.

Noggle is probably concerned about second-level controversies: Even if we all agreed that the *general* standards for the appropriateness of beliefs, desires, and emotions are those mentioned in his paper, their application would depend on evaluations and subsidiary criteria that would link them to specific situations. For example, even if we agree that beliefs should be "relevant to the situation at hand", we may not agree on which particular belief is relevant in a particular situation. Still, the crux remains that a norm-based account can tell us whether a particular act is manipulative only if we can refer to norms and ideals that are specific enough to decide whether a particular lever setting is ideal for a singular person in a particular situation. How should we deal with this problem? Noggle argues that the norms and ideals on which the manipulative character of an action depends must be those of the *agent* at the time of her action, and that we need not worry about their objective validity (Noggle 1996, pp. 47–48).

Noggle's argument goes like this: The relevant norms and ideals must be thought of either as *objective* ones, or as those of the *recipient*, or as those of the *agent*. He then dismisses the first, "objectivist" option as "not promising", presumably because he deems "it […] not at all clear that there is any such thing as an 'objective' standard for appropriate emotion" (Noggle 1996, p. 47):

"Reasonable persons often disagree: is anger called for or mere annoyance, shame or mere regret, hope or worry? Indeed, it is not at all clear that there is any such thing as an 'objective' standard for appropriate emotion or that there could be an algorithm for defining which information is relevant to which situation." (Noggle 1996, p. 47).

This is probably not meant to imply that *all* normative standards for ideal lever settings are equally contestable or subjective. Noggle seems open to the idea that the relevant standards fall into different categories, that some are indeed objective while others are not. However, he does not discuss options for accommodating the fact that some reasons for beliefs, desires, or emotions may be 'objective' while others are not. One reason may be that they would make the account more complicated than any of the three options Noggle considers.

The second option, which we might call "recipient-centered", would be to take the *recipient's* norms and ideals as decisive. Noggle finds this option unattractive as well. For it would imply that an agent, due to her lack of knowledge about the recipient's ideals, could never know whether she is manipulating her recipient, which seems implausible (Noggle 1996, p. 48).

Given the arguments against the first two options, only the third, "influencer-centered" account remains viable:

"[F]or the purposes of deciding whether someone acts manipulatively, we need to worry not about what the right path really is, or about what the victim thought it was, but about what the person doing the leading thought it was. And to do that, we will have to consult that person's beliefs about what the right path was." (Noggle 1996, p. 48).

Again, this does not commit influencer-centrism to any form of *substantive* meta-normative expressivism or subjectivism (although it does not exclude these views either). Its subjectivism is merely "methodological" (Werner 2022, p. 253). We need not worry about the validity of an influencer's beliefs *in order to identify manipulation*. Thus, the core idea can be expressed as follows:

An agent A manipulates a recipient R if and only if A intentionally influences R's attitudes (beliefs, desires, or emotions) so that they do *not* conform as closely as they otherwise

would to what *A* considers ideal for *R* at the time of *A's* influencing *R* – i.e. measured against the norms and ideals that *A* thinks should apply to *R's* beliefs, desires, and emotions.

This account is meant to have an additional advantage over the alternatives in that it reveals a conceptual parallel between lying and manipulating. For the standard by which a liar's statement is measured is also what the liar *believes* is true. The influencer-centered account implies that manipulativeness is relational in a specific sense: Whether an action is manipulative is relative to the norms and ideals of the influencer at the time of her action.

2 The Puzzle

Interestingly, Alexander Fischer and Christian Illies offer a different interpretation of Noggle's account. They assume that it essentially links the manipulativeness of an action to the norms and ideals *of the recipient.* Based on this (mis-)interpretation of Noggle's account as recipient-centered, they argue that it remains unclear whether the pertinent norms are those that the recipient *actually has* or those that the recipient *should have* (Fischer 2017, p. 60; Fischer and Illies 2018, p. 32).

While there seems to be no direct textual basis for interpreting Noggle's account as recipient-centered, it probably results from a genuine difficulty in understanding Noggle's position: It remains unclear whether, how, and to what extent an influencer is supposed to identify with her recipient in determining what beliefs, emotions and desires would be ideal *for* the recipient *from* the influencer's own perspective (similarly Turza 2023, p. 37 n. 55).

This is problematic. For in order to apply the influencer-centered criterion of manipulativeness, we must be able to identify those of an agent's evaluative beliefs that count as her beliefs about what is ideal *for the recipient*. After all, the influencer's relevant evaluative beliefs must concern "the *other person's* ideal condition" (Noggle 1996, p. 50 emphasis added). But can we be sufficiently selective about these beliefs while remaining strictly indifferent about their truth and even their plausibility? What if an influencer holds (or is convinced to hold) beliefs about what is "good for others" that in fact have nothing to do with their well-being – should we still accept them as relevant criteria?

3 Cases and Intuitions

To get some grip on the issue, let us start from Noggle's own example of the racist counselor. According to Noggle, "a racist who attempts to incite racial fears may not intend to move the other person away from what *he* – mistakenly – takes to be the other person's ideal condition, and so we cannot accuse him of acting manipulatively" (Noggle 1996, p. 50). The example is intended to show that:

"[e]ven if the influencer has a culpably false view of what is our ideal, the influence is not a manipulative action so long as it is sincere, that is, in accordance with what the influencer takes to be true, relevant, and appropriate." (Noggle 1996, p. 50).

In determining the manipulativeness of the influence, we (observers) need not care whether the agent's beliefs about the recipient's ideal condition are true or justified. It is enough to know that the agent 'sincerely' holds them. Still, we may wonder whether an

agent's beliefs need to satisfy any additional requirements in order to count as beliefs about what is ideal *for the recipient*.

To see the importance of this question, it may help to consider different variations on Noggle's example. It could involve a racist affected by "white fear" interacting with someone he considers his fellow. The agent would then be someone who *identifies with* and even, albeit in a misguided way, *cares about* this particular recipient.[3] One could also imagine a racist engaged in a very different attempt to instill "racial fears", however. Imagine him as a school principal talking to a student who belongs to a group whose members have been oppressed for generations and whom the principal considers inferior. By speaking in an elaborately condescending manner, he intends to deepen the student's insecurities and instill in her what he, the racist, considers "a healthy dose of racial fear": fear of what he sees as the power and superiority of his own group. The principal assumes that it is ideal for the student to feel this kind of fear, because it reminds her of what he considers her proper place in society.

This principal's action is utterly despicable. Should we also call it manipulative? According to influencer-centrism, we probably should not. After all, the principal is merely aligning the student's feelings with what he sincerely believes to be her ideal state. Of course, we could make up a more specific version of the story in which the principal, while believing it ideal for his student to be fearful, also holds ideals about reasons for fear according to which he considers his student's being addressed by his own condescending voice to be an inappropriate reason for the student to feel fear. But alternative versions of the story seem no less plausible. The principal may have no ideals about the appropriateness of reasons for fear, or he may consider them inapplicable to the student, or he may even regard his elaborate arrogance as an authentic expression of his own group's superiority and thus as a valid reason for the student to feel a mixture of awe and fear. Influencer-centrism would then classify his action as non-manipulative.

This seems implausible to me, but intuitions about the case may still be mixed. So let us tweak the example a bit more. Imagine that the principal wanted to not only instill fear, but also interfere with the student's ability to think clearly. His action would then match what Hanna sees as one of the "paradigm cases" of manipulation: "an agent uses a certain tone of voice because she knows that it will deter the target from thinking on her own" (Hanna 2015, p. 630). Nevertheless, depending on the principal's ideals, influencer-centrism may still not support the intuition that he is acting manipulatively. The principal may find it appropriate for the student to simply respond submissively and unthinkingly to his demands; he may find it ideal for her to be dominated by one of "his group" rather than to think for herself.

Should we accept the conclusion that even efforts to interfere with another person's cognitive functioning should not be considered manipulative if they are consistent with what the agent deems ideal for the target person? Perhaps we should first test our intuitions on a truly extreme example, one that is similar to Anne Barnhill's case of *Maria*, the "extreme narcissist" (Barnhill 2014, p. 67): Imagine an agent named *Egon* who deems nothing valuable but the satisfaction of his own desires. Unlike *Maria*, who has "no opinion about what is rationally and morally ideal for other people" (Barnhill 2014, p. 67), he is convinced that he knows what is ideal for other persons: Their beliefs, desires, and emotions must be in the state that best serves Egon's own interests. Therefore, he uses all means of influence –

[3] Even in this case, one may think it counter-intuitive to not call his inciting racial fears manipulative; cf. Jason Hanna's discussion of his "racist candidate" example (Hanna 2015, p. 633).

rational persuasion, coercion, deception, playing on other's emotions, etc. – that promise to further his goals. Since, in his view, the ideal state of others is whatever serves his purposes at the time, he can never have any reason to lead them astray from what he considers "their ideal condition". Thus, as long as he uses others to serve his own self-interest, Egon can never manipulate, according to (our reading of) the influencer-centered account.

In his reflections on childhood narcissism, Noggle actually seems to accept a similar conclusion. Observing that "children (and some adults as well) have an inflated sense of their own importance", he argues that when such "people […] sulk in order to get others to pay more attention to them" while they "actually believe they are entitled to that attention", they do "not in fact act manipulatively" (Noggle 1996, p. 50 cf. Note 17). On the other hand, Noggle describes manipulators as persons who treat their victims as if they "were some sort of object or machine" (Noggle 1996, p. 44). This is *exactly* what Egon does. After all, he treats others merely as instruments of his whims. Noggle may therefore be hesitant to extend his interpretation of the comparatively harmless case of sulking children and childish adults to all the potentially outlandish actions Egon would be willing to commit.

In any case, the assumption that Egon's selfish actions can never be called manipulative may seem puzzling. It also seems at odds with psychological terminology, as one of the reviewers of this journal thankfully pointed out. For Egon appears to exhibit character traits associated with the 'dark triad' (narcissism, machiavillism, psychopathy) which, in the words of the reviewer, are "routinely described as manipulative" (cf. Paulhus and Williams 2002; Furnham et al. 2013).

One might think that the last argument cuts both ways, though: First, if Egon is a clinical case, his example may not say much about the plausibility of influencer-centrism with respect to everyday cases. Second, given that Egon is a clinical case, one could deny his culpability. This would justify the conclusion that Egon cannot be a manipulator *if* we take culpability to be built into the concept of manipulation.

There is more to be said about both points, though. For unlike Barnhill's Maria, Egon is confident that he knows what is good for others, even though *we* deem his ideas off the mark. His case is therefore not *essentially* different from cases of sub-clinical egotism or even from the statistically normal cases of egocentric bias that frequently affects our judgments about what is good for others (the "impurity" of our hearts that Kant called the "principal affliction of human nature"; 2018, pp. 27, 53 f. [8:267, 6:29f.]). Egon's case is just the extreme end of a spectrum of egocentric bias in which the instrumentalization of others is based on – or rationalized by – self-serving misrepresentations of what is good for them. The challenge it poses to the plausibility of influencer-centrism applies to lesser cases as well.

In a nutshell, the problem is this: The more an agent's understanding of the target person's "ideal condition" is distorted by egocentric bias or limited by sheer lack of concern and diligence on the part of the agent, the less likely it is that influencer-centrism will call his self-serving interactions with others manipulative. This is counter-intuitive in a certain way. It seems natural to think that it should rather be the other way around. Egocentrism, lack of empathy, and lack of concern for others seem to be predictors of manipulative behavior and typical character traits of manipulators.

The point about culpability is more complicated. Egon's culpability could be denied for two different reasons. First, one might claim that he is not a responsible moral agent at all. This claim may be difficult to evaluate, but it may not be all that relevant. After all, if Egon is

just the extreme end of a spectrum, there are likely to be less extreme cases of persons who appear to be accountable agents even though their ideas about what is good for others are (sometimes) distorted by egocentric bias. Second, one might deny culpability for particular actions that reflect an influencer's distorted ideas about what is good for her recipients: If she does not distract them from what she (falsely) considers their ideal state, she lacks a guilty mind. This allows for two possible responses. First, one could argue that influencers do have some responsibility to overcome their egocentric biases, and that lack of diligence in this regard is actually one of the defining characteristics of manipulativeness.[4] Second, one could argue that acting manipulatively does not require a guilty mind (I will return to this briefly in the next section).

Even if one accepts Egon's example as a legitimate test case for assessing the plausibility of influencer-centrism, one might still bite the bullet and simply accept the implications of influencer-centrism. To get a sufficiently comprehensive picture of these implications, let us consider a final class of cases. To my knowledge, these cases have not previously been discussed as test cases for influencer-centrism. In my view, they provide the strongest evidence that the account is inconsistent with our intuitions. Interpreting these cases through the lens of influencer-centrism would produce 'false positives'. They would register as cases of manipulation, when intuitively they are not.

Imagine that Egon is not a textbook case of cold-hearted machiavellianism after all: He sometimes feels weak and unable to deceive or mislead others for his own advantage. Like real agents, he is a complex character and his actions do not always reflect his values. Sometimes he finds himself "in the grip of" (Gibbard 1990, p. 60) social norms that he does not truly endorse. Sometimes he experiences flares of natural empathy that he is unable to suppress. As a result, he sporadically acts honestly or kindly in ways that do *not* best serve his own goals (as he is painfully aware), but that critical observers would consider to be in the true interest of his recipients. But Egon being Egon, he is convinced that he is leading them astray from their ideal condition. Should we really call such actions *manipulative*, as influencer-centrism would suggest? As a less abstract example, imagine that the racist principal, still convinced that it would be good for his student to remain uneducated, is once prompted by professional habits to be more helpful, appreciative and informative in his interactions with her than he thinks appropriate. Should we really conclude, as influencer-centrism would have us do, that the principal *manipulates* his student by interacting with her in a non-intimidating, respectful way that promotes her personal growth?

This may be stretching the concept of manipulativeness too far. If our reconstruction of influencer-centrism is correct, it does not fit well with common uses of the term "manipulation" or "manipulative".

[4] This is very much in line with Marcia Baron's interpretation: "Manipulativeness […] involves arrogance, manifested in at least two ways: in her supposition that others' decisions are for her to make, and in the presumption (in the case of paternalistic manipulativeness) that she knows the other's needs, priorities, and weaknesses better than he does." (Baron 2003, p. 49).

4 Discussion

Nevertheless, there may be reasons to adopt influencer-centrism, understood as a *revisionary* account of manipulative action. It might help us to better understand or to improve (some, or some aspects, of) the practices in which questions of manipulation arise. Accordingly, this section will first examine two potential arguments in favor of influencer-centrism that have already been mentioned: the culpability argument and the alleged parallel between manipulation and lying. I will argue that neither of these arguments commits us to accepting influencer-centrism. The second part of the section will start from Noggle's concession that his account would make it hard to determine the manipulativeness of another person's behavior (Noggle 1996, p. 51). In what I believe to be a powerful new argument against influencer centrism, I will try to show that it is also unable to provide sufficient first-personal guidance for those who wish to avoid manipulating others: Its methodological subjectivism leaves its criteria of manipulativeness under-determined.

Concerning the culpability argument, two general points can be made. Some authors argue (Hanna 2015, p. 635 f.) or are at least sympathetic to the idea (Manne 2014) that we do not need to understand manipulation as implying culpability. Even taking the opposite view would not necessarily require us to accept influencer-centrism. A culpability requirement only requires us to understand manipulativeness as implying some sort of *mens rea*. According to influencer-centrism, however, a certain state of mind – an awareness that one is leading the target person astray – is not only a necessary but also a *sufficient* criterion for declaring the respective action manipulative. There is certainly room for accounts that combine some kind of guilty mind criterion with additional criteria of manipulativeness (e.g. Barnhill 2014).

A similar point can be made about the supposed parallel between lying and manipulating. Hanna (2015, p. 635) argues for abandoning the idea that the two concepts belong to the same family. If we want to keep it, however, we could try to do so by accepting the position that lying requires the objective falsity of a claim in addition to the speaker's belief that it is false.[5] Moreover, even within the framework of influencer-centrism, there remain differences between lying and manipulating, as I will point out at the end of this paper.

One limitation of influencer-centrism has been recognized from the beginning: It makes it nearly impossible for persons other than the manipulator to recognize manipulativeness. As Manne (2014, pp. 228–229) points out, this limitation is very similar to the limitation of recipient-centered accounts led Noggle to adopt influencer-centrism in the first place: Whereas recipient-based accounts would make it difficult for the agent to know whether she is manipulating, influencer-centered accounts limit the ability of others (including recipients) to know when an agent is manipulating (them).

Noggle acknowledges this point, stating that his account renders "it […] difficult in practice to determine whether someone acts manipulatively" (Noggle 1996, p. 51). However, he does not consider the problem where he first chooses between the three options, suggesting that he considers it less serious. The reason for this may be that limitations to the detectability of manipulativeness by recipients and observers 'only' affect the ability of others to know whether some manipulation took place – and hence their ability to assign "blame or

[5] Some accounts of lying require an untrue statement or assertion for there being a lie while others require only a speaker's own *belief* that her statement or assertion is untrue (Mahon 2016). Noggle adopts the latter view (Noggle 1996, p. 47).

punishment" (Noggle 1996, p. 51) – but not the ability of agents to avoid manipulating anyone in the first place. Assuming that it is more important to avoid instances of manipulation than to be able to blame manipulators after the fact, this would indeed justify a preference for influencer-centrism over recipient-centrism.

In a later paper, Noggle reevaluates the limitations of influencer-centrism and suggests a division of labor between influencer-centered and objectivist criteria of manipulativeness:[6] In determining whether an agent "was guilty of acting manipulatively" as a matter of "interpersonal ethics", we should adhere to influencer-centrism and refer to the agent's criteria of the recipient's ideal state. On the other hand, when deciding "public policy questions", such as "whether to encourage, discourage, or even prohibit healthcare professionals from employing nudges", we might better refer to objective criteria (Noggle 2018, p. 168).

Such division of labor, however, would not solve the problem entirely. First, it would risk tearing 'ethical' and 'policy'-uses of the concept apart, since it would not provide a systematic link between the influencer-relative 'ethical' and the objective 'policy-' criteria of manipulativeness. Moreover, since each individual agent would still remain the sovereign source of the criteria for determining the 'ethical' manipulativeness of her respective actions, the ability of observers to detect 'ethically' relevant instances of manipulation – and thus the ability of the moral community to maintain intersubjective practices of blame, justification, and excuse with respect to such instances –, would still remain precarious.

This brings us back to Noggle's original idea that "[b]ecause it is so difficult to tell when someone acts manipulatively" on the influencer-centered account, "we will have to try to get people to look into their own hearts and examine their own intentions" (Noggle 1996, p. 51). This suggests rather clearly that the real strength of influencer-centrism is not in governing intersubjective practices of "blame or punishment" but in governing the first-personal deliberation and self-criticism of agents who seek to avoid manipulating others.

What guidance does influencer-centrism offer agents who want to "look into their own hearts and examine their own intentions"? It tells them not to distract others from "the *agent's* conception of which beliefs, desires, and emotions are ideal *for the influenced person*" (Barnhill 2014, p. 66 emphasis in the original).

However, this guidance seems insufficient. To see why, let us take a short step back. Our discussion of examples like Egon or the racist principal suggested that taking even an agent's bizarre ideas about what is good for others as criteria of manipulativeness may have counter-intuitive implications. It also suggested that it may not even be clear which of an agent's beliefs should count as her notion of what is ideal *for the recipient*. For example, it may appear to a critical observer that the racist principal's beliefs about what is 'good for his student' are in fact much more about the principal's own longing for power and privilege than about *the student's* flourishing. In other words, his beliefs about what is good 'for' the student may not only be wrong, they may not even be beliefs *of the right kind*.

The problem for an account that remains indifferent about the validity the agent's evaluative beliefs is that the two questions (about the validity and about the object of his beliefs) are intertwined. We can only determine whether an agent's beliefs are really beliefs about what is good for the recipient when we take a stance about their merit, or at least their plausibility: Whether x is really good *for* P cannot be determined without evaluating whether x is really *good* for P – just like determining that a biography is one *of*, say, David Hume must at some point invoke criteria for *good* (or at least *decent and plausible*) biographies of David

[6] I am grateful to one of this journal's reviewers for suggesting to discuss this proposal.

Hume. Of course, we could just take the biographer's word for it, or blindly believe the book's title. But if nothing in the biography corresponds to the life of David Hume, while it resembles the life of Aretha Franklin, or the biographer's own youth, this would seem rather odd. Yet, without abandoning its indifference to the validity of an influencer's beliefs about what is good for the recipient, influencer-centrism seems committed to the same solution: It needs to accept whatever an agent sincerely declares or believes to be her notion about what is good for the recipient.

This leads to another problem that, to my knowledge, has not received sufficient attention: Uncertainty about which of an influencer's evaluative beliefs are those of the right kind does not only affect critical observers of social interaction. It also affects the first-personal deliberations of an influencer who wants to avoid manipulating others. She too may wonder which of her beliefs represent her convictions about what is ideal for her recipient. If she turns to influencer-centrism for help, she may feel that it places two demands on her that are are difficult to reconcile. For it tells her to be true to the ideals that *she herself* endorses (we may call this the "endorsement requirement"), but also to judge what is good *for the recipient* (we may call this the "identification requirement").

How should the influencer reconcile these requirements? There is a wide range of options. On one end of the spectrum, the influencer would simply ask herself what she would find ideal if she, otherwise unchanged, would find herself in the spacio-temporal location of the recipient. This clearly falls short of the identification requirement. Since the recipient may have different needs, commitments, and values, it may not tell her much about what is truly good *for the recipient*. On the other end of the spectrum, the influencer would ask what she would find ideal if she would share *all* of the recipient's traits, values, and commitments. But this would violate the endorsement requirement. The influencer's assessment would no longer depend on *her own* ideals but on (her interpretation of) those endorsed by the recipient. Influencer-centrism would then (almost) collapse into *recipient*-centrism. This possibility is probably what explains Fischer's and Illies's interpretation of Noggle's account (see Sect. 2).

So it seems that the endorsement requirement and the identification requirement pull in opposite directions. How could they be reconciled? Any influencer trying to avoid manipulating others would need to decide this question, but influencer-centrism cannot provide guidance. For to do so would require it to take a stance on what the recipient's ideal state really is. Thus influencer-centrism seems under-determined. While it suggests that influencers who try to avoid manipulating others should care about what is good for recipients, it does not tell us what this actually means.

5 Alternatives and Conclusion

If this is true, the limitations of influencer-centrism seem serious: As a reconstruction of ordinary uses of the term, influencer-centered accounts of manipulativeness seem implausible with respect to certain cases. As a framework for intersubjective practices of moral blame, excuse, and justification, its reliance on the standards of influencers limits our ability to determine whether an act is actually manipulative.[7] As a moral compass for first-

[7] It may also limit our ability to compare general types of manipulative behavior with regard to their moral significance, though more would have to be said about this point.

personal deliberation, influencer-centrism seems underdetermined and provides insufficient guidance.

What are the alternatives? Within the scope of this paper, I can only give a few hints.

The simplest response would be to fall back on another of Noggle's three options and try recipient-centrism or objectivism instead. But that may in fact be too simple.

Pure recipient-centrism seems implausible for at least two reasons. In addition to Noggle's objection that it would render it difficult for agents to know when they manipulate others (see Sect. 1), it would be blind to cases wherein a manipulator exploits pre-existing vulnerabilities like those of an irrationally guilt-ridden or submissive person (Hill 1991, p. 4 ff.).

Objectivism, on the other hand, has been invoked more than once as a means of improving or supplanting Noggle's 1996 account. As noted above, Noggle himself later proposed a division of labor between influencer-centrism and objectivism. His recent "mistake account" (Noggle 2020b) seems to avoid the issue: By arguing that manipulation is essentially an attempt to get recipients to make some kind of mistake, he seems to remain neutral on the question from who's perspective mistakes need to be identified. However, Noggle's insistence that the mistake account should be seen as an *extension* rather than a *revision* of his earlier account (Noggle 2020b, p. 250) may express an ongoing commitment to influencer-centrism. In any case, pure objectivism as defended by Hanna (2015) would face serious challenges as well. First, its implication that an agent could manipulate a recipient by moving her attitudes in a direction that they both sincerely believe to be ideal – perhaps even after serious examination and reflection – seems counter-intuitive. Second, Noggle's argument against objectivism still holds: There may be no "objective" standards for the appropriateness of *all kinds* of a recipient's relevant attitudes (Noggle 1996, p. 47).

If none of the three pure options is fully convincing, there may be something wrong with the trichotomy itself and/or with the underlying framework that presents it to us.

A comparatively modest revision would be to retain idea that manipulative acts essentially move the recipient's attitudes away from their ideal condition, but to develop a *hybrid* concept of that ideal condition which combines elements from more than one option. Barnhill, who amends influencer-centrism by a sophisticated reference to the objective self-interest of the recipient (Barnhill 2014, p. 72), seems to be a proponent of this approach. Hybrid accounts do seem more promising than pure ones with regard to reconstructing our linguistic and moral intuitions. They also complicate matters. Not only do they need to spell out criteria for objective concepts like "self-interest". They also have to tell a more complicated story about how and why the proposed combination of objective and subjective criteria makes sense from a moral point of view.

To some extent, this may seem fair and appropriate. Identifying manipulation may be more complicated than simple influencer-centrism suggests. We may have already suspected as much after our first attempt to interpret the notion of an agent's beliefs about her recipient's ideal condition. For any *plausible* interpretation would require of the agent to aim at some sort of *rational reconstruction* of her recipient's ideal condition in a way that is, among other things, (a) sensitive to the differences in situations and interests, as well as their respective interpretations, between the agent and the recipient, (b) sensitive to disagreements about ideal attitudes, (c) aware of her own epistemic limitations and vulnerabilities and sensitive to those of the recipient, (d) willing to engage with the recipient's reasons in an open evaluative discourse, aiming towards the goal of mutual understanding, and (e) aware

and respective of the spaces for non-rational forms of self-expression and commitments including existential choices.

This list is certainly sketchy and incomplete. But it may still support a general point: Caring about the ideal condition of others is a demanding task. The appeal of influencer-centrism lies not least in its apparent simplicity. But if something similar to the list is necessary in order to spell out what is required of agents who really care about acting non-manipulatively (and thus about other's ideal condition), then the simplicity of influencer-centrism is deceptive. The account seems simple in part because it is under-determined with respect to how influencers should determine the ideal condition of their recipients.

A structural feature of influencer-centrism may even make matters more complicated than they need to be. This may become clearer if we consider a difference between lying and (the influencer-centered view of) manipulating. Noggle suggests a strict analogy between the two:

„To lie is to attempt to bring it about that someone believes what is false; to act manipulatively is to attempt to bring it about that someone falls short of the ideals." (Noggle 1996, p. 48)[8]

However, he also seems to accept that "lying outright" (Noggle 1996, p. 45) requires the liar to make a statement that she believes to be false (this is what Mahon 2016 calls the "statement condition" of lying). This would mean that the prohibition of lying can be linked in a very straightforward way to the speaker's own epistemic commitments: It requires her to not contradict her own commitments. The case of manipulation seems to be much more complicated, because the relevant "ideals" of the influencer are not about her own epistemic, volitional and emotional state, but about that of other persons. This raises additional questions about the *methods* by which influencers may 'pull the levers' of their recipients even toward their ideal position.

To address these additional questions, the relevant "ideals" must include *procedural* ideals (e.g. about what would be an appropriate "mental process" of producing an emotion; Noggle 2020b, p. 250) in addition to substantive ideals (e.g. about what emotion would be appropriate in a given situation). Moreover, both kinds of ideals would have to be integrated in a way that is determinate enough to evaluate specific interactions.

In order for this to work, we must be able to derive a sufficiently comprehensive set of *social* norms for non-manipulative interactions *between* agent and recipient from norms for *the recipient's* ideal (or at least 'non-faulty'; Noggle 2020b) mental state and mental processes alone. Even if successful, this might not be the most straightforward and transparent approach. Notwithstanding their pronounced differences, alternatives such as Michael Klenk's *neglect account* (according to which manipulative influence is exerted with indifference towards whether the method of influence reveals pertinent reasons to the recipient; see Klenk 2022, p. 97), Christiane Turza's *functionally defective reasons* account (Turza 2023), and also *hidden influence accounts* such as those of Susser, Rössler, and Nissenbaum (Susser et al. 2019a, b) seem to have at least an advantage in terms of transparency. For they articulate at least some of the criteria of manipulativeness in the form of social norms that are directly applicable to the interaction between agents and recipients.

[8] Moreover, Noggle's discussion of Iago's case (Noggle 1996, p. 44 f.) suggests that he accepts the "statement condition" (Mahon 2016), though it is also not mentioned in this passage: "lying outright" (Noggle 1996, p. 45) requires the liar to make a statement that he believes to be false.

Acknowledgements I would like to Thank Jonas Dietz, Charlotte Gauckler, Tim Kirchner, Victoria Oertel, and Especially Birthe Frenzel and all Reviewers for many Valuable Suggestions and Comments on Earlier Versions of this Paper.

Author Contributions Micha H. Werner is the sole contributor to the article in its current form. I would like to thank several persons for helpful comments and suggestions, though, not least the reviewers of thisjournal. The final version of the manuscript contains a footnote with acknowledgements.

Funding Work on the article has been conducted as part of the regular activity of a professor of philosophy at a state-funded German university; no additional funding was received.
Open Access funding enabled and organized by Projekt DEAL.

Data Availability Apart from the literature mentioned in the "references"-section of the paper, which is publicly available, the submitted work does not draw on additional data and materials.

Declarations

Statement Regarding Research Involving Human Participants and/or Animals No research involving human participants and/or animals has been conducted in connection with the submitted article.

Ethical Approval Not applicable.

Informed Consent Not applicable.

Competing Interests There are no financial or non-financial interests that are directly or indirectly related to the article, apart from the author's interest in contributing to a philosophical discussion.

References

Barnhill A (2014) What is manipulation? In: Coons C, Weber M (eds) Manipulation: theory and practice. Oxford University Press, Oxford; New York, pp 51–72

Barnhill A (2016) I'd like to teach the World to think: Commercial Advertising and Manipulation. J Mark Behav 1:307–328. https://doi.org/10.1561/107.00000020

Barnhill A (2022) How Philosophy might contribute to the practical Ethics of Online Manipulation. The Philosophy of Online Manipulation. Routledge, New York, pp 49–71

Baron M (2003) Manipulativeness. Proc Addresses Am Philos Assoc 77:37–54. https://doi.org/10.2307/3219740

Bennett WL, Livingston S (eds) (2020) The Disinformation Age. Cambridge University Press, Cambridge

Buss S (2005) Valuing autonomy and respecting persons: Manipulation, Seduction, and the basis of Moral constraints. Ethics 115:195–235

Cave EM (2007) What's wrong with motive manipulation? Ethical Theory Moral Pract 10:129–144

Coons C, Weber M (2014a) Introduction. In: Coons C, Weber M (eds) Manipulation: theory and practice. Oxford University Press, Oxford; New York, pp 1–16

Coons C, Weber M (eds) (2014b) Manipulation: theory and practice. Oxford University Press, Oxford; New York

Fischer A (2017) Manipulation: Zur Theorie Und Ethik Einer Form Der Beeinflussung. Suhrkamp, Berlin

Fischer A, Illies C (2018) Modulated feelings: the pleasurable-ends-model of Manipulation. Philos Inq 6:25–44
Furnham A, Richards SC, Paulhus DL (2013) The Dark Triad of personality: a 10 year review. Soc Personal Psychol Compass 7:199–216. https://doi.org/10.1111/spc3.12018
Gibbard A (1990) Wise choices, apt feelings: a theory of normative Judgment. Harvard University Press, Harvard
Gorin M (2014a) Towards a theory of interpersonal manipulation. In: Coons C, Weber M (eds) Manipulation: theory and practice. Oxford University Press, Oxford; New York, pp 73–97
Gorin M (2014b) Do manipulators always threaten rationality? Am Philos Q 51:51–61
Gorin M (2022) Gamification, Manipulation, and domination. The Philosophy of Online Manipulation. Routledge
Hanna J (2015) Libertarian Paternalism, Manipulation, and the Shaping of preferences. Soc Theory Pract 41:618–643
Hill TE (1991) Autonomy and self-respect. Cambridge University Press, Cambridge; New York
Hudders L, Lou C (2022) The rosy world of influencer marketing? Its bright and dark sides, and future research recommendations. Int J Advert 0:1–11. https://doi.org/10.1080/02650487.2022.2137318
Jongepier F, Klenk M (2022) The Philosophy of Online Manipulation. Routledge, New York
Jongepier F, Wieland JW (2022) Microtargeting people as a Mere means. In: Jongepier F, Klenk M (eds) The Philosophy of Online Manipulation. Routledge, New York, pp 156–179
Kant I (2018) Kant: Religion within the boundaries of Mere reason: and other writings, 2nd edn. Cambridge University Press, Cambridge; New York
Klenk M (2020) Digital Well-Being and Manipulation Online. In: Burr C, Floridi L (eds) Ethics of Digital Well-Being: a Multidisciplinary Approach. Springer, pp 81–100
Klenk M (2022) Manipulation: sometimes hidden, always careless. Rev Soc Econ 80:85–105. https://doi.org/10.1080/00346764.2021.1894350
Mahon JE (2016) The definition of lying and deception. Stanf. Encycl. Philos
Manne K (2014) Non-machiavellian manipulation and the opacity of motive. In: Coons C, Weber M (eds) Manipulation: theory and practice. Oxford University Press, Oxford; New York, pp 221–245
Noggle R (1996) Manipulative actions: a conceptual and Moral Analysis. Am Philos Q 33:43–55
Noggle R (2018) Manipulation, salience, and nudges. Bioethics 32:164–170. https://doi.org/10.1111/bioe.12421
Noggle R (2020a) The Ethics of Manipulation. Stanf. Encycl. Philos
Noggle R (2020b) Pressure, trickery, and a unified account of Manipulation. Am Philos Q 57:241–252. https://doi.org/10.2307/48574436
Paulhus DL, Williams KM (2002) The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. J Res Personal 36:556–563. https://doi.org/10.1016/S0092-6566(02)00505-6
Sherman LE, Hernandez LM, Greenfield PM, Dapretto M (2018) What the brain 'Likes': neural correlates of providing feedback on Social Media. Soc Cognit Affect Neurosci 13:699–707. https://doi.org/10.1093/scan/nsy051
Sunstein CR (2016) The Ethics of Influence: government in the age of behavioral science. Cambridge University Press, Cambridge
Susser D, Rössler B, Nissenbaum HF (2019a) Online manipulation: hidden influences in a Digital World. Georget Law Technol Rev 4:1–45. https://doi.org/10.2139/ssrn.3306006
Susser D, Rössler B, Nissenbaum HF (2019b) Technology, autonomy, and manipulation. Internet Policy Rev 8. https://doi.org/10.14763/2019.2.1410
Thaler RH, Cass R (2021) Sunstein. Nudge: The Final Edition. Revised Edition. Penguin Books, New York
Turza C (2023) Manipulation: Zum Begriff und ethischen status. Brill mentis, Paderborn
Vrontis D, Makrides A, Christofi M, Thrassou A (2021) Social media influencer marketing: a systematic review, integrative framework and future research agenda. Int J Consum Stud 45:617–644. https://doi.org/10.1111/ijcs.12647
Werner MH (2022) Manipulation and the Value of Rational Agency. In: Horn C, Santos R dos (eds) Kant's Theory of Value. De Gruyter, pp 241–262
Zuiderveen Borgesius FJ, Möller J, Kruikemeier S et al (2018) Online political microtargeting: promises and threats for democracy. Utrecht Law Rev 14:82. https://doi.org/10.18352/ulr.420