Introduction and overview of chapters

Jongepier, Fleur; Klenk, M.B.O.T.

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

1 Introduction and overview of chapters

Fleur Jongepier and Michael Klenk

Nor from mine own weak merits will I draw
The smallest fear or doubt of her revolt,
For she had wifi and chose me.
                    – Revised passage from *Othello*, Act 3, Scene 3

1 Modern-day Iago

Shakespeare's Othello depicts a paradigmatic case of manipulation: Iago is jealous of Othello's relationship with Desdemona and forges a deceitful plan to tear them apart by making Othello believe – falsely – that Desdemona is cheating on him. Amongst other things, he places a handkerchief in the luggage of one of Othello's close confidants that Othello gave as a gift to Desdemona. Upon finding the handkerchief, Othello falls for Iago's trap and believes that he was betrayed by Desdemona. Iago's plan succeeds: a clear case of interpersonal manipulation.[1]

Interpersonal manipulation can also happen online. A modern-day Iago may have arranged for Othello to find misleading but suggestive messages on Desdemona's social media account to achieve the same effect. Or he may have harnessed more sophisticated technological means to manipulate messages exchanged between Othello and Desdemona through their voice assistant or smart fridge. And perhaps, there are new forms of interpersonal manipulation that an online modern-day Iago could realise, for example moderating and influencing what people see online and which content they are exposed to. Manipulation is as old as the history of mankind. And yet there are important reasons to be especially concerned about manipulation taking place online, in particular the scale and the nature of online manipulation. First, the scale: what is perhaps most striking about the online world is our increased interaction with algorithms and (autonomous) machines. One editor of this volume, for instance, has screen time warning pop-ups installed but happily clicks *Ignore warning for today* in order to continue scrolling on Twitter and Instagram. The other editor deleted emails from their phone but simply keeps logging back in through the browser. World-wide, people spend about two and a half hours on social media *every day*

(people in the Philippines winning – or losing – the match with a whopping three hours and 53 minutes).[2] Netflix has 72.9 million users on average, and YouTube almost 200 million, with 80% of US parents of children of ages 11 and below indicating that their kids watch YouTube.[3] Almost 40% of the US population uses voice assistants.[4] Most important of all, when it comes to scale, is the breaking down of the online/offline or real-life/digital life boundary, given that our lives are becoming increasingly immersed with (online) technologies.

Of course, the pervasiveness of technology in our daily lives and how technologically blended our lives have become is, as such, no reason to think manipulation must be everywhere, too. It is, however, a reason to be especially alert in light of the tremendous influence that technology seems to have on us. The modern-day Iago is not the CEO of Google or Alibaba per se; Iago may also be hiding in our smartwatch, our Wifi-controlled lights, our robot vacuum cleaner, our care and sex robots, our children's smart dolls, and our pets' remotely controlled food machine. So yes, manipulation has always been around, and we've known billboards and dubious salesmen for a long time. Right now, however, looking at how our interactions are shaped online, we appear to be dealing with salesmen on steroids and billboards that follow us around and that change depending on who's looking at them.

A second reason to be especially concerned with modern-day Iagos concerns the nature of online manipulation. Iago is a bad and cunning person, but at least we can understand, conceptually, his cunningness to some extent and have some sense, morally, how to evaluate his actions when his evil ways are brought to the surface. Human manipulation can be just as awful – perhaps even more awful – than technologically mediated manipulation, but we typically know, who manipulated us, and which moral–emotional responses would be (very roughly) appropriate.

All of this is very unclear when it comes to being manipulated by YouTube videos, voice assistants, personalised Google search results, Candy-Crush, political parties-using-Facebook, and so on. It is often unclear that we are manipulated. Online manipulation is rarely "brought to the surface." Whereas in Othello there is Emilia who, in the end, uncovers Iago's manipulation, there are not many online equivalents of Emilia in the digital age. The question of "who" manipulated us is even harder to answer, if that question makes sense at all. And rather than disappointment or anger that many of us experience in light of human manipulation, the typical moral–emotional response when one is subject to online manipulation is either confusion, a feeling of powerlessness, or simply indifference or fatigue ("ah, another scandal"). The type of agency and intentionality (not) exhibited by algorithms and more advanced online machines is complex and unclear, making societal–philosophical questions about their manipulative potential all the more acute. This volume aims to address these and other questions about the conceptual and moral nature of online manipulation. Here, we

will discuss the aim of this volume in some more detail and provide an overview of the chapters.

2 This volume

Behind the recent public and academic "techlash" seems to be the growing concern that the influence exerted on us by algorithmic systems or more advanced technological machines like robots can be distinctly *manipulative* and for that reason especially problematic. At the same time, the debate about online manipulation rests on philosophically vexed and, to some extent, underexplored territory. Philosophical attention to manipulation is luckily on the rise (see, for instance, Coons and Weber 2014), and scholars have begun to explore how manipulation differs (or not) from coercion, persuasion, nudging, and other forms of influence, as well as whether manipulation necessarily constitutes a moral wrong of some kind, and, if so, why. The existing literature is still relatively scarce, however, and when it comes to the literature on online manipulation, it is often simply stipulated or suggested, rather than argued for, that a certain technology or technological development is manipulative, and it is sometimes just assumed that *because* its manipulative it must therefore be morally problematic. However, manipulation might well in some cases be morally unproblematic or indeed desirable, so the inference from "manipulative" to "immoral" is not always evident. Also, various technologies, actions, or developments might turn out to be morally problematic not because they are manipulative but because they are coercive (say). Finally, it is not always clear whether some technological tool or online design would be immoral rather than merely (very) annoying for internet users.

All in all, many fundamental questions about both the nature of online manipulation and its normative status deserve more systematic attention. For instance, must online manipulation (always) involve "intentions" of some sort, and is such a thing as manipulation by a non-human agent possible? Is online manipulation necessarily opaque, or can one be manipulated online "out in the open"? As for questions in the normative domain, is online manipulation always morally wrong, and if so, why? Can online manipulation also be morally acceptable or even a morally good thing to do? Does being manipulated online threaten autonomy, and if so, what do we take autonomy to be?

This edited volume aims to fill a critical gap in current discussions regarding the conceptual nature and moral status of online manipulation. We aim to provide theoretical and normative depth and nuance to debates in digital ethics about the manipulative influence of algorithms and autonomous systems. Thereby, we aim not only to enrich "applied" debates about online manipulation by bringing in contemporary developments from the philosophical debate regarding manipulation but importantly to also enrich and

sharpen the philosophical debate by putting existing theories to the test by applying them to online cases and contexts. Finally, we hope to make a methodological contribution by offering a type of applied philosophy that is solidly anchored in philosophical theory whilst strictly in the service of contributing to contemporary societal questions and challenges.

3 Overview of chapters

This volume is the first to explicitly address the philosophy of online manipulation. It contains 20 previously unpublished chapters and brings together leading international philosophers and several promising scholars at earlier stages in their careers. We sought to illuminate the questions surrounding online manipulation specifically from a perspective informed by moral and political philosophy. The chapters in this volume fall under the following four parts:

Part I: Conceptual and methodological questions
Part II: Threats to autonomy, freedom, and meaning in life
Part III: Epistemic, affective, and political harms and risks
Part IV: Legal and regulatory perspectives

Any ordering of contemporary contributions to novel philosophical and societal developments is bound to be artificial to some extent, and this volume is no different in that regard, as most authors cover more than one, and sometimes all, of the aforementioned broader themes. Still, it is possible to observe differences in emphasis and focus. For instance, contributions falling under the first heading are primarily concerned with the conceptual question of what manipulation is, how we should go about defining the notion, and how (if at all) online manipulation is different from offline manipulation. Chapters falling under the second heading are principally concerned with the moral dimension of manipulation, addressing the question of what, if anything, would make online manipulation immoral, and what exactly is at stake or threatened when a person is manipulated online, with a specific focus on threats to autonomy, freedom, and meaning in life. Contributors clustered under the third header consider possible threats to knowledge, control of our emotions, and political legitimacy. Finally, a separate heading is reserved for contributions that zoom in on a specific technology (such as real-time profiling) and then go on to ask how, for that technology, regulation is currently arranged and how it might be improved.

3.1 Part I: Conceptual and methodological questions

In the opening chapter, titled "Online manipulation: charting the field," we – the editors – present an overview of what we consider to be some of the core questions surrounding the nature and normative dimension of offline

and online manipulation. Our aim is not to settle these questions once and for all but to provide an overview of the theoretical landscape so that the reader is in a better position to locate and appreciate what is at stake in the other chapters that follow. We touch upon some methodological and conceptual preliminaries and then give a brief overview of so-called outcome- and process-based accounts of manipulation, noting their advantages and disadvantages. In the second part of the chapter, we consider what we call "aggravating factors" that help explain the distinct problems raised by manipulative online technologies, such as personalisation and opacity.

In the opening chapter, we mention quite a number of philosophical controversies and nuances regarding the conceptual nature of manipulation. Indeed, many discussions about manipulation, online or offline, involve asking the question "Are these kinds of influence *actually* instances of manipulation?" However, in chapter 3, "How philosophy might contribute to the practical ethics of online manipulation," Anne Barnhill argues that asking that question might not be the most productive way for philosophy to contribute to the debate and that we should be careful not to get bogged down in philosophical definitions and demarcation issues. Instead, she suggests that when online influence is called "manipulative," we should try to figure out what kinds of *concerns* are being registered by calling it manipulative and then query whether influence of that particular form is problematic and why.

In Chapter 4, Massimiliano L. Cappuccio, Constantine Sandis, and Austin Wyatt turn to the very distinction between online and offline manipulation in their chapter "Online manipulation and agential risk." They ask how manipulation enabled by AI-based technology that mediates our interactions online (such as recommender systems on social media) differs from other forms of manipulation. The authors draw on developments in communication science to suggest that different technologies enable different "communication paradigms" which, in turn, engender different forms of manipulation. They then turn to what they refer to as "engagement-maximization-based online manipulation" and argue that this is best thought of as an emergent phenomenon, not traceable to the explicit or implicit intentions of any individual agent but more akin to collective action.

The next two chapters address the very possibility of speaking sensibly about online manipulation or manipulation by machines. In Chapter 5, titled "Manipulative machines," Jessica Pepp, Rachel Sterken, Matthew McKeever, and Eliot Michaelson ask how the contemporary concept of manipulation could capture current and future instances of manipulation by machines. They provide a clear overview of the different theoretical positions one could take and introduce helpful insights from the conceptual engineering literature. They suggest that one might use the concept of manipulation as if machines could manipulate us, even if they don't literally do so. And they present an ameliorative approach which involves asking which *purpose* is served by having a certain concept and also allowing to

change our concept of manipulation in order to make better sense of, and make room for, genuine machine manipulation.

In Chapter 6 "Manipulation, injustice, and technology," Michael Klenk defends a specific proposal about manipulation by technology. Understanding technology quite broadly, he shows that it has considerable effects on us independently of whether it is (artificially-) intelligent, autonomous, or embodied. He argues that being manipulated should be understood differently than manipulating. On his account, a manipulated mental state is one that is explained in the relevant way by an injustice. Drawing on considerations about epistemic injustice and the affordances of technology, he argues that technology can contribute to injustices that explain our mental states in relevant ways. Therefore, we can be manipulated by technology, independently of whether technology has, for example, intention.

3.2 Part II: Threats to autonomy, freedom, and meaning in life

When it comes to making online choices, an oft-heard concern is that these choices are manipulated and therefore not autonomous. In Chapter 7 "Commercial online choice architecture: when roads are paved with bad intentions," Thomas Nys and Bart Engelen turn to the question of what exactly is manipulative about commercial online choice architectures (COCAs) and in what way they threaten personal autonomy. They argue that considering the intentions of the manipulator is key, both conceptually and normatively speaking. They end their chapter by pointing out that even in cases where the intentions of internet users and COCA designers happen to align, there is still cause for concern as the latter are typically completely indifferent towards the aims of the former.

Fleur Jongepier and Jan Willem Wieland pick up the thread relating to indifference in Chapter 8 "Microtargeting people as a mere means." In this chapter, Jongepier and Wieland focus on political microtargeting and propose that what is wrong about employing such techniques is that they involve treating people as a mere means, which they argue involves genuinely caring about people's consent to be used in certain ways. They go on to explain what "caring about consent" comes down to in digital contexts and argue that political microtargeting typically, though not necessarily, involves treating people as a mere means due to a lack of care about people's consent to be used as a means towards the microtargeter's ends.

Next, Marianna Capasso argues in Chapter 9 "Manipulation as digital invasion: a neo-republican approach" that neo-republicanism can provide conceptual and normative tools to analyse and address the problem of manipulation in relation to digital nudges. The neo-republican approach offers a promising account of the connection between digital choice architecture and human freedom given its emphasis on social and political relations as well as collective and shared responsibility. Capasso individuates specific criteria to assess when digital nudges can amount to dominating

manipulative interferences or "invasions." She argues that the main worry about digital nudges is not (just) the fact that they are typically not transparent but that it involves alien control and a lack of democratic means of empowerment, communication, and contestation.

In Chapter 10 "Gamification, manipulation, and domination," Moti Gorin remains within the Republican framework and focuses specifically on gamification, that is, the attempt to turn an activity into a game, to make it fun, engaging, and motivating. One of the examples of online gamification discussed by Gorin is Twitter, whose system of likes, retweets, and so on can be seen as introducing the so-called game reasons into human discourse, where such reasons would not ordinarily exist. Gamification turns out to be manipulation on Gorin's account because it is a kind of influence that makes people do something for game reasons rather than any other reasons that may ordinarily exist. Based on this analysis, Gorin presents an analysis of the wrong-making features of manipulation inspired by Republican worries about domination and offers an account of domination which he calls "interactive domination" that differs from the structural domination articulated by republican theories.

W. Jared Parmer likewise focuses on gamification in Chapter 11 "Manipulative design through gamification." Parmer focuses on gamification as it offers a useful starting point for understanding manipulative design more generally. Gamification is the implementation of inducements to 'striving play' for the sake of purposes beyond those typically found in games, such as to learn a skill or to develop certain habits. According to Parmer, gamification becomes manipulative when it involves deception, on the part of the manipulator, about her purposes. Parmer points out that one of the dangers about manipulative design is that it stands in the way of making our lives more meaningful because it can make it harder to work out and act on what we care about..

The relation of manipulation and meaning in life brought out by Parmer nicely connects with Chapter 12, "Technological manipulation and threats to meaning in life," by Sven Nyholm. Nyholm first offers a helpful overview of the different positions that one may take on the question of whether technology can manipulate humans. He then turns to the more general question regarding the relation of manipulation and meaning in life and provides an overview of different constituents or contributors to a meaningful life. Nyholm then argues that technological manipulation threatens some or all of these factors, thus endangering the opportunities of those interacting with the technology to enjoy meaning in life. Nyholm's chapter contributes to a better understanding of the normative dimension of manipulation as it suggests that it is a type of influence the effects of which are particularly harmful.

Geoff Keeling and Christopher Burr then consider the question of what distinguishes morally permissible from morally impermissible behavioural influencing strategies by software agents. They argue that morally

impermissible instances of behavioural influence by software agents undermine the "mental integrity" of human users. In other words, such strategies diminish people's capacity for authentic decision-making. Such strategies, they argue, are morally permissible only if behavioural influence by software agents affords due respect to the mental integrity of the user.

3.3 Part III: Epistemic, affective, and political harms and risks

Within the focus on normative and evaluative aspects of manipulation, we then shift perspective to consider which epistemic, affective, and political harms and risks may be associated with manipulation.

In Chapter 14 "Is there a duty to disclose epistemic risk?" Hanna Kiri Gunn focuses on personalisation of online platforms and in particular on the epistemic risks involved. In many online spaces, she argues, we risk undermining our ability to be in reasonable control of our epistemic capacities, for instance through the personalisation of search engine results or being exploited by bots to spread fake news or emotional persuasion. Gunn argues that internet users are placed at risk of social-epistemic harm without their informed consent and that there is a moral duty to disclose the social-epistemic risks of using online services to prospective users. She closes the chapter by zooming in on moral responsibility and the many hands problem.

Lukas Schwengerer is likewise concerned with the epistemic dimension in Chapter 15 "Promoting vices: designing the web for manipulation." He is primarily concerned with normative and evaluative questions surrounding the problem with manipulation, which he approaches through a discussion of user-friendly design. Schwengerer takes an innovative virtue epistemic perspective to suggest that user-friendly design promotes an "overly trusting attitude" towards the information provided by the website. Schwengerer argues that artefacts like websites can warrant trust to a given degree. Trusting them beyond that degree "destroys the virtue of intellectual carefulness." When we lack that virtue, we are easier targets for manipulation because we might more readily and less critically believe, feel, or desire what the website's creator wants us to believe, feel, or desire. The virtue epistemic perspective makes it easy to see why that would be bad, and it is interesting in the context of our volume for making explicit the link between epistemic vices and potential for manipulation.

Next up are two chapters that deal, in different ways, with the link between online manipulation and emotions. Nathan Wildman, Natascha Rietdijk, and Alfred Archer focus on "Affective online manipulation" or the online influence on people's affective states. They begin by considering four key questions to distinguish different types of manipulation, such as whether it is active or passive, done intentionally or unintentionally, based on a top-down or bottom-up mechanism, and finally whether the aim is primarily to influence affective states or, ultimately, behaviour. Their next step is to consider why any of this would constitute manipulation.

They consider three prominent accounts and suggest, in a pragmatic vein, that each of them can account for the manipulativeness of online affective influence, albeit in different ways. The authors argue that in extreme cases, online affective manipulation constitutes a distinct type of injustice, namely "affective powerlessness," in which someone (or something) wields a large amount of power over the emotional states of the user, rendering the user affectively powerless.

The focus on the affective component of manipulation is continued in Chapter 17, "Manipulation and the affective realm of social media," by Alexander Fischer. He focuses on both the nature of manipulation and its moral evaluation. Fischer argues that manipulation manifests itself in changing the victim's evaluation of a given end as pleasurable or displeasurable. Hence, unlike coercion, which may force a given end upon the victim, manipulation merely moderates the attractiveness of an end and thus its likelihood to be chosen. In the second part of the chapter, Fischer turns to social media, and he gives several examples and cases to illustrate how social media impacts our affective states, thus making it a powerful tool for manipulation.

In Chapter 18, "Social media, emergent manipulation, and political legitimacy," Adam Pham, Alan Rubel, and Clinton Castro begin by observing that political advertising and disinformation campaigns on social media can have a significant effect on democratic politics. Pham, Rubel, and Castro point out that often the moral concerns with these activities are reduced to the effects they have on individuals, such as the fact that their autonomy is undermined. The authors instead suggest, by introducing and analysing the concept of "emergent manipulation," that the presence of manipulation in electoral politics threatens the legitimacy of the elections themselves, and thus that the wrongness of such activities is to be found at the group level.

3.4 Part IV: Legal and regulatory perspectives

Kalle Grill's chapter, "Regulating online defaults," concerns the normative aspects of manipulation, which he explores through a discussion of online defaults and how they may be regulated. A default option is an option from which one can only opt out by taking an action. Grill shows how online defaults – which have become inevitable features of online environments – can distract, misinform, harm, and eventually manipulate people. Grill's second main contribution is to consider principles for the regulation of defaults, including that they should be set to favour non-consumption, that data collection is minimised, and "that information provided by default is true, or at least not demonstrably false or against expert consensus."

In the final chapter of the volume, Jiahong Chen and Lucas Miotto discuss the morality of real-time profiling, that is, the collection of information about an individual's present status to generate a profile in an attempt to influence the individual's actions in the immediate future based on that

profile. Zooming in on real-time profiling, they argue, allows us to see what is morally problematic with manipulation more generally. The authors argue that real-time profiling is morally wrong because it involves "psychological hijacking" and because, by making the user more vulnerable, it makes them more likely to be wronged in other ways too. The authors then turn to regulatory measures and discuss the implications for consumer protection law and data protection law and their limitations, arguing that a more targeted regulatory approach is needed to effectively address the unique challenges of real-time profiling.

4 General observations and concluding remarks

The contributions in this volume span across a wide spectrum, not just in terms of how conceptually or normatively oriented they are but also in terms of the technologies the authors focus on and the methodologies they (explicitly or implicitly) use. It should not be surprising this volume as a whole would not give us the true theory of online manipulation and why it is or isn't morally problematic. More than anything else, the chapters taken together give the reader a clear view of the state of the art when it comes to the philosophy of online manipulation. This view is bound to be kaleidoscopic because it includes philosophers who are very much concerned with getting the philosophical definition of "manipulation" right before moving on to the "online" adjective (whereas others get right to it); philosophers who are very much concerned with threats to individual persons (and others much more with threats to the collective, social, or political order); and so on. In other words, this volume will not give the reader "the" approach to studying online manipulation. However, it will, we hope, give a rich, kaleidoscopic view of many of the concepts, methodologies, moral concerns, and applications that are at stake in this debate that has only started to unfold.

When we consider all the chapters taken together, a few observations can be made. First, it is interesting to see how many chapters in this volume do not just "employ" concepts and theories from the philosophical debate about (offline) manipulation but really – as we, as editors, hoped – also challenge and test these theories by applying them to the online sphere. Second, it is interesting to see that many (though not all) contributors in the volume do not have a detailed and settled position on what they take manipulation to be, what exactly sets it apart from persuasion or coercion, whether it is necessarily opaque or intentional, why it's wrong, and so on. This can be indicative of the fact that both the philosophical debate about manipulation and the debate about online manipulation are still very much in development and there is as yet no clear "map" on which to position oneself. Also, it might be indicative of an (implicit) pragmatic methodological approach (to be discussed in the next chapter), namely that it is possible to

have illuminating discussions about various aspects of online manipulation without necessarily providing a fine-grained definition of manipulation first.

Third, across all chapters the terms "online" and "technology" really stand in for a wide range of phenomena. We have discussions that understand "online" or "tech" in terms of highly general design approaches such as user-friendly design, default-settings, or gamification which are applicable to all technological designs. Others discuss more specific affordances of recent online and algorithmic technology such as social media, real-time profiling, and augmented many-to-many communication. Each individual contribution makes clear what the relevant factors are that may be seen as aggravating the problem of manipulation.

The fourth and final observation. In the original call for chapters for this volume, we were operating with a distinct "conceptual" and a "normative" part for the prospective book. As it turns out, this two-part ordering of the book did not make much sense in the end. Even though a couple of authors are clearly more concerned with either the conceptual side of online manipulation or with the normative side, by far most of the contributors really have an equal interest in both. In other words, we could say that to answer normative questions about why certain forms of online manipulation would be problematic in some way, one inevitably needs to enter some theoretical terrain (if only briefly). The converse is also true: to make progress on the question of what online manipulation is, conceptually speaking, it is hard if not impossible to say something about the instances in which it is (or appears to be) morally problematic. This, on its turn, may tell us something about whether or not "online manipulation" is a so-called thick concept (which is something discussed in the subsequent chapter).

It is important to point out some of the limits of this volume. Though this book, with its many chapters and diverse approaches, is very comprehensive, many other questions remain to be addressed and answered. For instance, this volume is heavy on the (moral and theoretical) theory and relatively light on the "what now?" question. Two chapters explicitly address regulation and policy issues, and many other authors also briefly discuss what the practical consequences of their account might be. Still, the emphasis is more on *understanding* online manipulation and applying new and existing philosophical resources to do so. Second, even though some authors make use of material from other disciplines (law, social sciences, and so on), this is not an interdisciplinary volume on online manipulation. It is a philosophy-based book on online manipulation, which has the aim of making certain developments in philosophical debates relevant to (as well as testing them against) developments and technologies in the online world. Despite it not being an interdisciplinary volume on the subject, we of course do very much hope that it will – by bringing in a lot of (sometimes neglected) philosophy – be of use to scholars from other disciplines working on online manipulation and related topics. Taken together, if there were going to be a second volume

or follow-up to this book, it would take an interdisciplinary approach from the get-go, and it would be heavier on the "what now?" side.

We hope that this volume will help us and others to continue the discussion and motivate and inspire further work on this societally acute and philosophically intriguing topic.

Notes

1. Both authors contributed equally to this chapter. Fleur Jongepier's work was supported by a VENI grant. Michael Klenk's work was supported by the European Research Council under the Horizon 2020 programme under grant agreement 788321.
2. https://techjury.net/blog/time-spent-on-social-media/
3. https://techjury.net/blog/netflix-statistics/
4. www.emarketer.com/content/voice-assistant-and-smart-speaker-users-2020