The University Of Sheffield.

Manipulation and Influence: A Trickery Account of Manipulation Applied to Three Scopes

James Dunstan

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Arts and Humanities
Department of Philosophy

Submission Date

30/04/2023

Table of Contents

Acknowledgements 3

Abstract 4

Declaration 5

Introduction 6

Chapter One: Manipulation as Influence 8

- Influence 8
- Manipulation as bypassing or subverting reason 11
  - Discussion of Mills' Version of Subverting Reason 15
- Manipulation as Pressure 16
- Summary 19

Chapter Two: A Definition of Manipulation 21

- Noggle's Definition of Manipulation as Trickery 21
  - Critical Analysis of the Manipulation as Trickery Account 23
  - Manipulation as Trickery and Intention 28
  - Manipulation and Bullshit 31
- Problems and Replies 33
  - The Sincere Manipulator Problem 33
  - The Odd Manipulator Problem 36
  - Manipulation, Self Interest and Paternalistic Manipulation 39
  - Covert Manipulation, Agent Authorship and Temptation 41
- Summary 45

Chapter Three: The Ethics of Manipulation 47

- The Wrong-Making Features 48
- The Moral Status of Manipulation 51
  - Baron and Strawson 51
  - Manipulation and Small Wrongs 57
  - Conclusion 59
- The Moral Responsibilities of the Manipulated 59
  - Summary 67

Section Two 69

Chapter Four: Mass Manipulation 69

- Introduction: 69
- The Undefined Audience Problem 71
- The Undefined Manipulator Problem 74
- Acting Collectively 76

Artwork, Media and Manipulation 80
Conclusion 84
Chapter Five: Manipulation and Design 86
Introduction: 86
Nudge Theory: What is a Nudge? 86
Dark Patterns and Dark Systems: Applying the Trickery Account 93
Conclusion 101
Reference List 103

Acknowledgements

I should first give my thanks to Dr James Lenman, my primary supervisor for this thesis, and Dr Chris Bennett, my secondary supervisor. My deepest thanks go to both for the thoughtful and challenging conversations, sage guidance and the open support for my ideas and the project. Whilst I might have been able to complete it on my own, given enough time and luck, it would likely be in a sorry state indeed without the learned advice of both of you. I should also thank all the staff at the Department of Philosophy at the University of Sheffield, both academic and administrative. I've been studying at Sheffield my entire academic career, and I cannot imagine having studied anywhere else. Every staff member, lecturer, and peer during my time at Sheffield has contributed in some way to my development and enabled me to write this thesis. I'd also like to thank friends and family who have supported me over the years, kept me going when things seemed impossible and gave their thoughts where they could. Special thanks are also needed for my partner Mollie Hodgkinson, without whom I would have given up a long time ago. I want to dedicate this work to her, as well as to my parents, Susan Dunstan and John Dunstan. Without them this work would not exist.

Abstract

Manipulation is a form of influence that has received little attention compared to other forms of influence, such as coercion. Those accounts of manipulation which do exist often fail to address how a single account can function within different scope. An account should be able to explain manipulation in the context of an abusive relationship just as well as it explains mass manipulation between groups such as corporations and nations towards large, non-specific audiences. My thesis aims to correct this by developing Robert Noggle's trickery account. I will explain and solve some key flaws in the account and also argue the definition of manipulation should be extended to include manipulation by 'bullshit', following from Harry Frankfurt's concept and his use of it regarding deception, which the trickery account necessarily counts as manipulation. Next, I explore the ethics of the act of manipulating, the manipulator and potential responsibility of agents acting under the influence of manipulation. Finally, armed with a robust, developed account of manipulation as trickery, I explore how the theory can be applied to account for mass manipulation to large audiences, as well as how individuals and groups can manipulate through the design of systems and interfaces. I argue these scopes each provide unique challenges which must be solved by any account. Overall, I argue that the trickery account, with further development, offers a complete account of manipulation.

Declaration

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not previously been presented for an award at this, or any other, university.

Introduction

If there is one type of influence that we're the most concerned about at the moment, it might be manipulation. Nobody wants to think of themselves as being manipulated. It's a concept seeped in literary examples of evil individuals treating those around them as disposable tools. Nobody wants to be the disposable tool of somebody else's grand plan. Examples of manipulation, both imaginary and real, range from grand, nation-wide schemes to the mundane mind games of lovers in a toxic relationship. With the rise of the internet, smartphones and so on, it almost goes without saying that there are now more opportunities and avenues for some agents to influence other agents. Some of this influence we might be concerned about, for a whole host of reasons. However, at least some of our concern intuitively rest on worries about whether these technologies, and the people behind them, are manipulating us in some way.

Consequently, just what makes an act manipulative is something important to know. Sadly, it is a difficult task to define just what it means to manipulate. The main problem is that the term as used in common language can refer to a whole range of behaviours which can seem very different from one another. This makes it very difficult to find a single definition which is not *underinclusive*, in that it does not define as manipulative some behaviours we intuitively think are manipulative. At the same time, we do not want to make a definition so broad that it becomes *overinclusive*, including behaviour we don't think of as manipulative. Much of this also requires we consider how we view the concept of manipulation ethically. Manipulation as a term is a pejorative, meaning it has negative connotations. However, it would be unwise to assume that our common intuitions hold true, and that manipulation is ethically wrong. We will need to examine in detail how manipulation works, and what wrong making features, if any, it possesses.

Currently, most scholars have done either one or two of these tasks. Much has been written about what a definition of manipulation should include, or a definition given in relation to one context or another only. Others have looked to what's wrong with manipulation if we assume a definition, without interrogating whether that definition should be adopted. There are few attempts to produce a full account manipulation, which provide both a definition of manipulative acts and take an ethical stance on manipulation as they have defined it.

Yet these accounts are limited in that they face valid criticisms which as of yet nobody has adequately resolved. More crucially to my mind, they fail to acknowledge the true scope any definition of manipulation faces. Most accounts, understandably, focus on examples of manipulation actions which are, for a lack of a better term, intimate. They are between two lovers, family members or work colleagues. This is useful to isolate an examine how a manipulative act functions on this small scale. However, it is this sort of manipulation, which while concerning, we are currently culturally the least concerned about. Rather, we are most concerned about politicians who appear to be manipulating large swathes of the population, technologies which appear to place the power to manipulate our children in the hands of a small few companies or countries etc. A definition of manipulation should work on every scale, from the intimate to the multi-national, if it is to be useful to us in our current context.

In this thesis, I have two tasks. The first is to look at available accounts of manipulation and find a candidate for further development. This means solving any flaws it has, and developing it into a robust, complete account. The second task is to then take this account and expand its scope, examining cases of what I will call mass manipulation and manipulation through design. By the end of this thesis, I hope to not only have put forward a strong case for why the manipulation as trickery account, as I will develop it, is a unifying account of manipulation that we should adopt, but also

demonstrate that it is an account which can apply to all examples of manipulation, both small and large.

In order to do this, the thesis will be split into two parts, and five chapters. The first chapter will consider manipulation's place as a method of influence, conceptually taking its place alongside persuasion and coercion as categories of influence. I will then in this chapter run through the development of accounts of manipulation in literature and identify which one is the best candidate to be further developed. In chapter two and chapter three, I will proceed to provide a definition of manipulation and an account of its ethics respectively. In the second part I will apply our definition to both mass manipulation and manipulation by design. The change of scope will produce unique challenges and opportunities for application to other areas of philosophical interest. For example, in the final chapter, I will examine first and foremost the 'manipulation objection' to nudge theory.

Chapter One: Manipulation as Influence

Influence

We are constantly being influenced. This has always been the case, but our modern world has increased the magnitude and salience of sources of influence. We are influenced when we see an advertisement on the television. We are influenced when we see a sign in a carpark warning us to lock our cars and take our valuables with us, lest they be stolen. We are influenced to some extent by everybody we have contact with, whether they be our families, friends, co-workers or perfect strangers. With the rise of smart-phones and social media, new channels of influence have been opened. Each buzz and ring of a smart-device prompts us to engage with technologies designed to influence us, for better or ill. Right now, as you read the ideas I have expressed, I am influencing you, whether because you are engaged or simply because I am boring you. In a manner of speaking, any interaction with the world is an influence of sort, in that you are affected and changed by the world as you interact with it. Experiencing influence is an inevitable consequence of experience.

To be influenced simply means to be affected by something. What is more interesting is how other people affect us, specifically in a social context. When I use the word influence in this context and going forward, I will be referring to *social influence*. To give a simplistic definition of social influence: An agent *X* influences another agent, *Y*, when *X* interaction with *Y* in a way which affects *Y*'s behaviour, judgments, desires, or emotions. Moreover, we can distinguish between intentional influence and unintentional influence. For example, a handsome man walking down the street may influence the people who see him. His looks may make some feel envious or attracted to him, and some may even act differently because of him. His presence has a social impact, but it's not intentional in the same way that speaking to someone with the aim of changing their behaviour is intentional influence.

To intentionally influence somebody implies the agent has a goal in mind. This would be a change the influencer wants to incite, usually with the further goal of making a person *act* how the influencer desires, as a consequence of whatever change is made. This change could also be reinforcement. To reinforce a belief is still to change it, only not in content. It is just held more strongly. Of course, the same would be true for other effects, such as making a person doubt their belief. If we wanted to consider whether a particular act of influence was ethical, we may consider the goal as the first obvious place to look. Consider if I persuade somebody to murder another person. In this case I have clearly done something unethical by influencing them. The extent to which I am responsible, compared to the person who acts, is a topic open for discussion. Yet we clearly view it as unethical. A person who hires a hitman, influencing them by giving them money for their service, clearly does wrong, and the law accords with this fact by making such influence a crime. On its own this fact is unremarkable, an act which intentionally causes something to happen which is unethical, is also unethical. What is more interesting is that we also have moral intuitions about the *method* of influence used, no matter our intuitions about the effect of the influence.

Typically, we break down the general concept of social influence into two methods (or types) of influence, persuading and coercing. Whether such methods of influence can be employed unintentionally, or only intentionally is a matter to be discussed in the context of each method of influence. For our purposes, we should understand that social influence is split into two methods that are relatively well defined. At least in so much as there is a breadth of available and comprehensive definitions for them. However, we may also intuitively think of another method,

manipulation. It occupies a sort of middle ground between the two. Take the following example to illustrate the types.

Bed Time: Jonathon is trying to get his son, Timmy to go to bed. Timmy does not want to go to bed. Jonathon first tries to persuade Timmy by telling him that if he doesn't go to bed, he'll be really tired the next morning, and will have a terrible day at school. Timmy is not persuaded by this argument. Jonathon then tries to manipulate Timmy. He tells him that if he goes to bed, Jonathon will give him ice cream for breakfast. Jonathon is lying to Timmy; he will not give Timmy ice cream for breakfast. Timmy still doesn't want to go to bed. Finally, Jonathon threatens that if Timmy does not go to bed right this instant, he will ban him from watching television for a week. Timmy goes to bed.

Jonathon goes through the three methods of influence here. Persuasion, manipulation, and coercion. If we are lied to, we may describe ourselves as feeling manipulated. If Timmy woke up the next morning expecting his ice cream, having dutifully gone to bed when this was offered, he would feel quite betrayed when he was served something different. Perhaps not all lies are manipulative, but this one appears does appear that way.

Manipulation does not have as long a literature analysing it as the other methods of influence. Yet, it is just as interesting and becoming increasingly relevant. We are concerned about how we are influenced, not just the direct consequences of the influence. As I have said, and wish to emphasise, we are forever being influenced by others, and in a world of technologies such as social media, we are being influenced more strongly than ever before. What is so interesting about manipulation is the sheer breadth of different examples we intuitively identify as potentially manipulative. Here are some examples of what can be called manipulative influence:

Apocalyptic Preacher: A television minister preaches that the end times will soon be upon us. On the same show, she advertises various emergency food rations, shelters and weaponry. She warns her flock that the prophesised end of days could happen any day and reminds them of the current offers on equipment to survive it. She has been doing this for the last ten years, no end of days has yet come. She is quite wealthy.

Jealous Lovers: Megan and Darcy are dating. Megan feels that Darcy hasn't been paying enough attention to her recently. When they are out shopping, Megan flirts with the shop assistant with Darcy present. Megan makes eye contact with Darcy and winks. Darcy knows Megan is only flirting with the assistant to make him jealous, yet he feels jealous anyway.

Conditioning by the Party: Smith is a child in a totalitarian regime. Like other children in his nation, at school he is forced to recite slogans espousing nationalist beliefs and support of the ruling party. Smith grows up to feel pride in his nation and with a desire to support the party.

Ticket Deception: Mr Fisk is an important businessman. He abuses the ticket salesperson at the train station when the machine which prints the tickets malfunctions. The salesperson purposefully tells Mr Fisk that his train is leaving from a different platform that it actually is. Mr Fisk gets on the wrong train and ends up late for an important business deal.

Peer Pressure: Sarah is trying to cut down on her drinking. Her friend Percy thinks she is being a stick in the mud and is a lot less fun when she is sober. Percy purposefully buys Sarah drinks when they next go out, and if Sarah tries to decline, he accuses her of being ungrateful for the free drink. Sarah ends up getting drunk.

Internet Slideshow: When browsing the internet, Jerry finds an article promising to tell him what happened to a notable figure from his childhood. Jerry clicks on the article and is directed to a

slideshow, which requires him to load a new page and go through fifty different pages before he finds the one fact he wanted to view. By Jerry doing so, the website earnt a great deal of money from advertising revenue than they would have done otherwise, as Jerry viewed fifty of their pages, all filled with advertisements.

Cold Bureaucracy: A government wants to reduce the amount of money it spends on welfare payments for its poor citizens, without receiving the bad publicity they would if they stopped providing some payments, or directly excluded some citizens from receiving them. Therefore, they make it complicated and difficult for people to apply for the payments, by asking for documents which most poor persons do not have and cannot afford. The process is so complicated that many fewer citizens receive the payments.

These situations are quite different from each other. They differ regarding the techniques which are used, the material consequences of the influence, and in the scale on which the influencer operates. In this thesis, my goal is to develop a definition of manipulation which can account for all these vastly different scenarios. One key criticism of some definitions of manipulation in the current literature is that they do not capture all of the scenarios we may intuitively think are manipulative. Of course, the extent to which this *is* a criticism is debatable, as it may be that when we develop a full definition of manipulation, we find out some of these intuitions, especially those not strongly held, must be cast aside. Inevitably, in clarifying what is and is not manipulative, we will come to conclusions contrary to some of our intuitions. This is a simple consequence of finding clarity, you validate or invalidate some intuitions. I should note here then that the examples above are intuitively manipulative to me, first and foremost. I do not argue you should share them here, nor that they are actually examples of manipulation. That would be putting the cart before the horse, so to speak. Though of course I would not include them if I did not intend to argue for them.

One thought to consider is why it matters that we identify some influence as manipulative, some as coercive, etc. As I have previously stated, the most obvious place to look when evaluating whether an act of influence is ethical or not, is the goal of the influence. It does not matter how you influence a person to commit murder, we might think. Rather we can assume that the act of influencing somebody to commit murder is wrong because *murder* is wrong. Turn this on its head, however, and we can also say it does not matter how you influence a person to do something good, such influence is ethical because what you are trying to influence them to do is good. This is much more intuitively problematic position. To force a person at gunpoint to give their money to a homeless man is theft, while to persuade him to part with his money is encouraging charity. The method of influence matters when considering whether the influence is ethical or not. We want to consider whether it is right not only for us to influence, but to influence in a specific way. In other words, the ends do not always justify the means.

That which motivates a study of influence is considering to what extent our choice of method of influence affects when, if ever, it is ethical to influence. To do this, we should be able to define and identify what sorts of methods acts of influence employ, and then ask questions of whether the method has ethical implications. For manipulation, the most common intuition is that to manipulate somebody is to rob them of their freedom in some way to choose what they want to do. We do not like being manipulated, being influenced without our consent or awareness. We do not like to be treated as machines, where pressing the correct buttons and pulling the right levers will produce the result the manipulator wants. To produce a full account of manipulation is to first identify what methods of influence we refer to as manipulation, then to conduct an ethical analysis of it.

To provide a complete and full account for the vast variety of scenarios where manipulation apparently occurs is challenging enough. However, many current accounts do not also consider the different actors involved in manipulation. There are three key scales of influence I say manipulation can occur:

1. **Direct agent to agent manipulation**. By this I mean influence which occurs directly between two or more people. Typical examples would be manipulative interactions between parents and children, romantic partners, or salespeople and their customers.
2. **Mass-manipulation**. Politics and advertising are the two areas where persons are regularly subject to influence they label as manipulative. A politician may deliver a speech to millions, of whom she has no knowledge, and manipulate them. An advertisement may be a symbol or image which effects people around the globe, when the person who designed it is no longer involved. They may even have died. Influence we call manipulation occurs on large scales as well as small ones, with different mechanisms of communication, and no clear manipulator.
3. **Manipulation by design**. Many people would label a casino as manipulative; in that it is designed to encourage gambling. Shops and supermarkets are designed so we navigate them in particular ways, such that companies will pay to have their products placed in particular places. The same is said of advertisements online. The design of systems we interact with, whether they be physical or digital, or a mixture of the two, have the vast capability to affect our thoughts and behaviours. Supermarkets, online or otherwise, are merely one example. This is a form of influence similar but distinct in my view from mass-manipulation. An account of manipulation should explain or constructively rule out this form of influence and how it might be manipulative.

With these preliminaries out of the way, we should see the true scale of the challenge ahead of us. I have not even begun to list the menagerie of questions which we might ask of an account of manipulation, and expect it to help us answer. The best place to start is with a review of the literature and trains of thought others have produced, so we can find what approach has the most potential to rise to these challenges and provide the best definition of manipulation.

Manipulation as bypassing or subverting reason

The first place many have started when considering manipulation, is by contrasting it with persuasion. Persuasion, loosely defined, is a method of influence which involves appealing to a person's reason. If a friend is considering moving abroad due to a new job opportunity, and we do not want them to move so far away from us, we would act to persuade them by providing them with reasons to stay. What sort of reasons might we give them? Perhaps we argue that it will be so expensive to move abroad that it will not be worth the cost. Perhaps we argue they will not like the climate in the country, or the culture. Perhaps we state we do not want them to move because we will miss spending so much time with them, and this reason is presented as a reason not to move and give up our close friendship. Whatever reason we give them, we are appealing to their capacity to reason. There are other examples of persuasion operating at different scales. A televised debate between two scholars on the existence of God involves both parties giving each other reasons which support their perspective and challenging the reasons the other scholar provides. Since it is televised, they are also giving us, the audience, reasons as well.

Manipulation, the line of thought continues, does the opposite of this. Manipulation does not engage with our capacity to reason. The manipulator does not want to work with us, they want to work against us, move us against our will or without our awareness to suit their own ends. The most

illustrative example which motivates this line of thought is the hypnotist. The hypnotist begins to swing their pocket watch and we are rendered helpless. We enter a trance and when we are released from it, we believe new things, or find desires we had now muted. Perhaps we even acted but were not aware of it, lost in the trance the hypnotist prompted. The example of hypnotism takes manipulation quite literally. We are manipulated in the same way a puppeteer would with a marionette. Another closely related 'archetypical' example is the subliminal advertising. This advertising is implicit. Symbols or messages are placed in media which is not otherwise overtly designed to advertise. The goal of the subliminal messenger is that our unconscious minds will recognise the messages, while our conscious minds remain none the wiser. We find ourselves with a craving for a particular product in the case of subliminal advertising, with no knowledge of why.

The efficacy of these techniques is somewhat overstated in popular culture. However, we can treat them as what they are, hyperbolic examples of what we intuitively identify as manipulative influence. For a more realistic example, we could consider some hypothetical examples of advertisements. Though hypothetical, I'm sure you will have seen an advertisement like these examples at some point or another.

Consider an advert which presents dramatic images of rich, powerful and attractive people living a life of prestige and luxury. If you hadn't seen one of these types of adverts before, you might find yourself wondering what product it could possibly be advertising. Finally, as it ends, the model sprays herself with a perfume, and the brand name of the perfume appears on screen. This advert and those like it, present no rational reason to buy the product. We may go further and say that the advert actually appeals to non-rational features in the audience, such as a desire for approval from their peers or a sense of insecurity about their current lifestyle. The advertisers want the audience to associate the perfume and its brand with prestige and economic status through juxtaposition of those images with the perfume. Of course, the perfume may smell pleasant, and wanting to smell pleasant is a perfectly good reason to buy the perfume. But I would contend that the advert itself provides no serious rational argument to purchase the perfume. The influence is entirely irrational. Another example may be a fast-food advert which simply presents mouth-watering images of the food you can buy, ending with their logo and brand name. A person may watch this and feel hungry, and now have on their mind a place where they can buy something to sate the craving the advert itself produced. Yet I wouldn't argue that such adverts are making any rational argument to the viewer as to why you should eat at their restaurant over others, or anything similar. They are just raising awareness of their brand over others, which does not require presenting reasons to the viewer.

This line of thought provides the starting point for defining manipulation. Manipulation is, in essence, the opposite of persuasion. One example of this is definition is that of scholar Cass Sunstein, who provides a definition of manipulation in his paper '*Fifty Shades of Manipulation*'.

"an effort to influence people's choices counts as manipulative to the extent that it does not sufficiently engage or appeal to their capacity for reflection and deliberation."[1]

He further explains how manipulation on this account functions using the general psychological distinction that agents have two systems at work when acting. System 1 is the non-rational side. This is the side of us that is mainly built on instinct and habit. It is our desire for food when we are hungry, our reactions of fear or shock to a loud noise, or the anger which causes us to lash out without thinking. System 2 is our rational, calculating capabilities. This is the side which evaluates

[1] (Sunstein, 2015, p. 6)

the facts about a situation in order to make a decision.[2] These rudimentary psychological concepts are useful because we can use them to explain what manipulation as bypassing reason actually does. The idea is that when a manipulator uses influence intending to bypass reason, they may choose to appeal exclusively to System 1. The advertiser of perfume is clearly appealing directly to the first system. The viewer may still use System 2 when watching advert. For example, we may be struck with a desire to buy the perfume, then consider whether we have enough money to afford it. The point is that the *influencer* is not appealing to System 2 at all, they intend to bypass reason entirely.

Clearly not all manipulation aims to bypass reason. Lies are the most obvious example. In **Ticket Deception**, the salesperson gives Mr Fisk a reason to go to the platform and take the train that arrives there. It just so happens that the train went to a different destination than the salesperson said it did. The salesperson relied on Mr Fisk accepting the reason he was given and trusting the salesperson. Lies which intent to prompt action rely on a person understanding the falsehood expressed and acting upon it as if it were true. If we want to accept that lies such as that in **Ticket Deception** are manipulative, we cannot look to bypassing reason as the only definition. We might say then that manipulation occurs when a person bypasses the reasoning capacity or *subverts* it. We now need to know what exactly it means to subvert a person's rational capacities.

Claudia Mills provides the most straightforward explanation of what it means to subvert the rational process in an agent in her paper '*Politics and Manipulation*':

"We might say, then, that manipulation in some way purports to be offering good reasons, when in fact it does not. A manipulator tries to change another's beliefs and desires by offering her bad reasons, disguised as good, or faulty arguments, disguised as sound—where the manipulator himself knows these to be bad reasons and faulty arguments. A manipulator judges reasons and arguments not by their quality, but by their efficacy."[3]

At least one of the purposes of our rational capacities, we may think, is to seek the truth.[4] We take in information, process it, and use that information as a basis for deciding and justifying our actions. If we are fed false information, or poor reasons, our rational capacities still function and are engaged but provide ineffective results. Our rational capacities are intentionally led to serve a different purpose that they are meant to, or that we want them to, ergo they are subverted. When the salesperson deceives Mr Fisk about the train he needs to take, Mr Fisk's rational capacities are engaged but fed false information. Mr Fisk's rational capacities work as intended but produce a false result.

We can produce a disjunctive definition by combining Mills and Sunstein:

Manipulation 1: A manipulates B when A influences B and either i) fails to sufficiently engage or appeal to B's capacity for reflection and deliberation; or ii) engages with B's capacity for reflection and deliberation, in such a way as to subvert the purpose of this capacity.

This definition explains **Ticket Deception**, but we may find it harder to apply to our other examples, at least not without further exploring and specifying the method of influence used in those examples. In **Conditioning by the Party**, we may have to assume that Smith is exposed to bad reasons for trusting the party as well as the slogans. Otherwise, it may be that while Smith is forced to recite slogans and this influences him, the content of those slogans may be counted as

[2] (Sunstein, 2015, p. 13)
[3] (Mills, 1995, p. 100)
[4] Perhaps more accurately, we seek the truth because it allows us to make better decisions to achieve our end. I am not committed to finding the truth being the absolute end of our cognitive function.

'sufficiently' engaging with his rational capacities. We might also consider if the influence is still manipulative if all the slogans are actually truthful. Similarly in **Peer Pressure**, we would have to assume that Sarah's rational capacities are being subverted by being provided with the temptation of the free drink, and additionally believe that the reasons Percy provide arguing she should drink (her being less fun to be around when not drinking) are not good ones, and thus subvert her capacity. We should also consider 'cherry picking'. A person might present only good reasons but be aware of and not present other reasons against a course of action. In **Conditioning by the Party**, it could be that the slogans are good reasons, but the party vehemently suppresses negative aspects of the party from its followers. Smith is presented good reasons, but all the other good reasons which would paint the party in a bad light never reach him. We might think this is an example of manipulative action, but the definition does not include it. Smith is sufficiently engaged and only engaged with good reasons. It is just that he is purposefully made ignorant by the party of other reasons.

With these concerns in mind, it should be clear that while we can work with this definition, the terms being used need much more explanation. It is not just enough to say we must sufficiently engage with an agent's capacity for reason, we must explain just what conditions must attain for it to be sufficient. We should not just say that one's rational capacities are subverted. Rather we should explain what subversion really means, what methods can be used and what different means there may be of subversion. The existing accounts of manipulation appear to follow the path of subverting or bypassing reason but vary wildly in their explanations of what these terms mean, and the processes thereby employed by manipulators. Thus, this definition, while a great starting point, must be filled in with more details to be of any use to us.

It is worth at the offset to consider the applicability of this definition to the examples I have provided which take emotion as the central means by which the influence occurs. Take **Jealous Lovers**, where Megan makes Darcy feel jealous by flirting with a shop assistant. Her desire is to make Darcy pay attention to her by making him feel jealous. One train of thought is to say that emotion is separate from reason. Consequently, a rational decision is one made separate from emotion. For example, if I am angry at a person and strike them, I may later think that this was an irrational decision. While yes, I was angry, striking them may have isolated me from my friends and stopped me from mending bridges with the person I have struck. Not to mention the fact that I may face legal troubles for my action. While emotions are an inescapable part of being human, one understanding of rational is that it is limited to reasons, arguments, and logic, not desires and emotions. To use the previous simple psychological model for decision-making, emotions belong to System 1.

To dismiss emotions as some sort of defect in rational decision making, an interfering force which harms our judgments, is a hyper-rationalist approach. While it has a large tradition, the separation between emotions and reason is arguably over-stated. At the very least, we can evaluate our emotions. We speak of somebody being 'irrationally' angry if they have become extremely upset over what we would think is a minor slight against them. A person who is terrified of spiders, no matter how small and insignificant a spider might be, may think of their fear as irrational. This person may think to themselves that they 'should not' be scared. In contrast, a person who sees a spider they know has a deadly bite in their bed would justifiably be terrified. I will explore this in more detail in the next chapter. At this juncture it suffices to say that for the purposes of defining manipulation, it will not do for all appeals to emotion to count as manipulative. An influencer may be appealing to the emotional system of an agent when, for example, they make them feel guilty for something they have done which has harmed the influencer. Influencing a person to feel guilty, and thus motivate them not to repeat the action, must be distinguished from making a person feel guilty

where the target of the influence has done nothing wrong. We should be able to distinguish between guilt which is warranted considering the context, and what is colloquially called a 'guilt trip', making a person feel guilty when there is no justified reason for them to feel so.

The subverting or bypassing reason approach is a promising starting point for a definition of manipulation, but many details need to be filled in to account for the variety of examples of influence we consider intuitively manipulative. Before moving to another approach, it is worthwhile to examine a few defects with the definitions provided by Mills and Sunstein.

Discussion of Mills' Version of Subverting Reason

To restate, Mills' definition of manipulative influence is that manipulative influence occurs when an agent purports to provide good reasons, but actually provides bad reasons for a belief, desire or act. The context she presents this explanation is in the political sphere. Clearly politicians often engage in this sort of influence. Policy makers and pundits can discuss the facts of a politician's platform. They can point out that a policy, such a climate policy, could have an insignificant or negative effect on society, contra the claims of the politician. Yet, if the politician consistently provides their reasons, steadfastly asserting they are good reasons for enacting the policy, many will believe them regardless of the facts of the matter.

Mills provides a list of campaign strategies she views as manipulative and I think that while we could quibble on the details, most would agree they are manipulative. It still behoves us to list a couple of the more interesting examples. One is fear-mongering, where a candidate for office exaggerates threats of economic decline if they are not elected and their rival's economic plans come to pass.[5] While it would not be manipulative to assert that the economic plans of the rival would be bad for the nation, we may assume there is no reason to believe the economy would collapse entirely. If the politician then engages in this over-exaggeration, they provide bad reasons to the public, but portray them as good reasons. Therefore, they act manipulatively. Another example would be partial truths. A politician might assert that their government passed a law on some issue which is now popular, but they lobbied against the bill at the time.[6] Without the full context, this is a misleading reason, and ergo a poor one. Nevertheless, the politician asserts it as a good reason to vote for them.

There are two problems with this definition which I do not think it can be adapted to solve without radically changing it. Firstly, it is unclear whether or not it is the giving of bad reasons themselves or the belief of the influencer that the reasons they are giving are bad ones which makes the act manipulative. She asserts that the manipulator must believe that the reasons they are offering are bad ones, but this leaves open the question of whether a prospective influencer attempting to give bad reasons, but actually giving good ones is acting manipulatively or not. This may be an open question which Mills could answer. We could allow that a person can intend to manipulate and attempt to. Yet they fail to actually manipulate.

The second problem is that we might think manipulation can occur with the presentation of only good reasons, not bad ones. Consider the following example of two colleague, Bill and Ben. Both work for the same company. Bill finds out that both him and Ben are candidates for promotion in the next month. Bill influences Ben by presenting him with reasons he should quit his current role. However, he does so only using good reasons Ben has to quit. Ben accepts these good reasons and quits the company, meaning Bill gets the promotion. Bill never mentioned to Ben that it was possible Ben would soon be promoted.

[5] (Mills, 1995, p. 109)
[6] (Mills, 1995, p. 109)

It intuitively seems that Bill manipulated Ben, but he did so only by providing good reasons (e.g., now is a good time to re-train, the management style at the workplace is poor, Ben has a long commute to work, etc). We might think Bill acted manipulatively because he concealed the important reason for Ben not to quit his job, that Ben might get promoted. However, Mills does discuss whether we have an obligation to present *all* relevant reasons for and against an action, and concludes aside from some special circumstances, we do not. This is reasonable, as it is an undue burden to have to present all sides to a discussion in most circumstances. It is fine for people to be bias and advocate for specific positions, without also having to explore reasons against their position. We might consider that if Bill and Ben were friends, rather than just co-workers, they may have a special relationship which means Bill does have an obligation to disclose this information. However, this would not seemingly change the content of the influence, only our evaluation of it as we are now considering other relevant moral concerns. It should be noted this is a problem which rears its head repeatedly when considering any definition of manipulation as bypassing or subverting reason.

It suffices to say that defining subversion as having something to do with giving people *bad* reasons is correct. However, Mills' definition can be developed, and I argue it would need to be so done in a way which expands on what makes a reason a bad one, as well as the importance of the intention of the influencer. Finally, it would need to be altered to include case we intuitively identify as manipulative where only apparently good reasons are presented.

It is precisely an account of manipulation which does these things, taking the bypassing and subversion approach to manipulation while solving these problems, which I will introduce in the next chapter and proceed to discuss and develop. However, for the moment I would like still to consider the place of other approaches.

Manipulation as Pressure

There is a separate starting point to defining manipulation which compares it more closely to coercion, rather than interference or avoidance of the target's rational capacities. The manipulation as pressure train of thought can categorise manipulation as a subcategory of coercion as a weaker form of it. Alternatively, manipulation as pressure is considered its own method of influence, but one which operates quite similarly to coercion. That is not to say that this meant to capture all example of manipulation. Rather manipulation as pressure definitions become part of disjunctive definitions or define only a single *type* of manipulation.

We can think of persuasion and coercion as occupying opposite sides of a spectrum, with manipulation falling in the middle. This spectrum tracks how able the person is to act freely. Persuasion, we might think, presents reasons and arguments to a person for them to evaluate. The decision they arrive at once they have considered the arguments was made by them and them alone, with the assistance of the arguments and reasons provided to them. Coercion on the other hand dramatically restricts the freedom of the person subject to it. Allen Wood, writing on manipulation as pressure writes:

"I am coerced to do something when I either do not choose to do it or if, when I do choose to do it, I do it because I have no acceptable alternative."[7]

Consider the archetypical case of coercion. You are travelling along a ride when a highway robber jumps out from some bushes, brandishes a pistol, and says: your money or your life! The intention of

[7] (Wood, 2014, p. 92)

the robber here is to place pressure upon you by way of a threat, such that you give him your valuables. Of course, you are still free to make a choice. Assuming his threat is sincere, if you were feeling quite suicidal you could disobey him so that you get shot. If you were feeling quite heroic, you might try to snatch the gun or otherwise fight him off. In these cases, you have failed to be coerced. The pressure was not enough to force your choice. For most people however, there is a choice between giving the robber your valuables, or risking being killed. The latter is unacceptable, simply meaning that it is not a state of affairs you are willing to let occur. Therefore, there is only one suitable choice remaining for you, acquiescing to the robber's demands. This has ethical implications. If a bank worker helps an evildoer rob a bank by letting them into the building at night, and they do so because they believed the evildoer had their child at gunpoint and was threatening to kill them, we might think the bank worker morally blameless for their actions. Of course, coercion is more complicated than this simplistic picture on most accounts[8], but it serves our purpose. Coercion is a sort of influence which aims to alter a person's choice situation, by making some choices they have unacceptable. The goal is to influence the person to pick the choice or choices which the influencer desires they choose.

Manipulation also puts pressure on the target, altering the choice situation. However, rather than making it unacceptable, it merely makes it less acceptable. In this sense *A* is manipulated by *B* when *A* intentionally alters *B*'s choice situation in a way which makes a choice *B* has less acceptable, but does not render it entirely unacceptable to *B*.

Peer Pressure is one of our examples where this sort of definition can explain what is happening. When Percy buys Sarah a drink she doesn't want, as she is trying to drink less alcohol for her heath, she tries to decline it. However, Percy exerts pressure upon her by acting upset with her for declining his generous offer and being a 'stick in the mud'. Crucially, Sarah can still choose to decline the drink, and deal with Percy acting negatively towards her, perhaps even complaining about her to their mutual friends. However, this is now a negative consequence Percy has introduced to the choice of not drinking. It does not render it unacceptable, but less preferable. On a grander scale, we might consider **Cold Bureaucracy** to also be accounted for using this definition. When the government makes the process for applying to receive certain benefit payments harder for its citizens, it does not render it impossible. A citizen may still embark upon the difficult process to receive the payment. However, the government has altered the citizen's choice situation, the choice to apply is more arduous, and therefore less appealing. The citizen has been pressured not to apply for the benefit.

The pressure need not be negative. We can imagine positive incentives being used to pressure a person. However, this requires an element of care. If I do not normally desire to paint your fence, but you offer me £100 to do so, then I have not been placed under pressure by your incentive. You have simply hired me to paint your fence. On the other hand, if I were extremely impoverished, and you offered me money to perform a humiliating or sexual act, then we could conceive that this is positive pressure which is manipulative. Where the line is to be drawn between manipulative incentive and the non-manipulative one is a difficult question. We might think that it is an offer which renders the choice of *not* choosing it unacceptable. However, we can still have positive incentives which do not seem manipulative using this standard. For example, a job might be offered to a person which is so well compensated that it is unacceptable to them *not* to take the position. Perhaps it must also be the case that the choice being altered would otherwise be unacceptable to

[8] See Nozick's *Coercion* for an example of how complicated the picture of a simple coercive act like the robber can be.

us, or we are in circumstances where we are particularly vulnerable to the incentive. For example, a drug addict with no money to fund their habit may find that their drug dealer will give them drugs if the addict agrees to sell drugs to others. The addict is desperate and only agrees because of their compulsion, they are then then pressured by the dealer to sell the drugs in a manipulative way.

One question regarding manipulation as pressure is how to define 'acceptable'. If I threaten that I will pour your mug of tea down the drain if you do not do as I say, then I alter your choice situation. However, the pressure is very slight. You would certainly not be coerced by this threat, but would you be manipulated if this was sufficient to alter your behaviour? I am inclined to say that you would be. While a very slight threat, in the context of some abusive relationships I could see that this minor threat would be sufficient to prompt arguments and feelings of guilt and powerlessness. In this case it would not be the loss of the tea itself which alters the choice, but the associated negative emotions. However, to many this is so inconsequential that it would not really be a threat at all. On the other hand, we should consider what changes to the choice situation are unacceptable, so as to draw a line between coercive and manipulative influence. We should consider that the idea of unacceptableness is subjective. For somebody very vain, the notion that they might receive a small scar on their face might be absolutely unacceptable, so a threat of this sort would be coercive. On the other hand, somebody who already has facial scars or facial tattoos might see this as a less significant consequence. If we return then to the example of a threat to pour your tea down the drain, for some people we can assume this really would be a significant threat, and thus they can be manipulated or coerced with this threat. It follows that the effective coercer will tailor their threat to be unacceptable to the intended target, based on what they know of the target. The same is true for the effective manipulator, who knows what pressure they can bring to bear such that their target is more likely to choose what the manipulator wishes.

Another question regarding manipulation as pressure is how it co-exists with definitions of manipulation as bypassing or subverting reason. Anne Barnhill provides a distinction between what she calls "manipulation by ideal response" and "manipulation by non-ideal response". In the former, the manipulator aims to change the target's choice situation, such that the 'ideal' response the target chooses is—or is more likely to be—a choice the manipulator desires. With the latter, the manipulator interferes in some way with the decision-making process of the target, causing them to make a non-ideal response, which is what the manipulator desires.[9] This can also occur by changing the situation in a way which causes the interference in the decision-making process.

A relatively benign and common version of manipulation by ideal response is 'badgering'. A child wants a new toy. They try and influence their father to buy it for them. They try and persuade them, by pointing out how cool the toy is, but their father is unswayed. However, the child keeps repeating their arguments, day in and day out, until finally their father just buys them the toy. By pressuring their father, the child altered their father's choice situation by making buying the toy the better choice, simply because it would shut the child up. It is also true to say that not all manipulation by ideal response is also manipulation by pressure. If I remove a choice from somebody, I also change their choice situation. You have a choice between three types of pizza, I eat one, I have influenced you to pick one of the other pizzas still available. This is altering your choice situation, but it hasn't affected the desirability of a choice in the way manipulation as pressure does.

These two means of influence are both referred to as manipulation, but clearly function differently. This is not to say that a person cannot engage in both at the same time. For example, consider workers whose job is to hand out flyers for events. Eager to hand out as many flyers as possible

[9] (Barnhill, 2014, p. 52)

within a short time span, workers may be rather aggressive in brandishing the flyers at passers-by. Consider one of these workers offers you a flyer, and you decline. They say something along the lines of 'come on, just take it so I can get paid'. Assume for a moment that this, or something like it, constitutes a guilt-trip. Their intention is for you to be made unreasonably guilty about denying them a quicker payday. This would be manipulation by non-ideal response. You have no rational reason to take the flyer, you'd just put it in the nearest bin anyway. Yet you feel guilty and take it. However, you are also being influenced by means of a rational response. Pressure is being put on you by the continued interaction. You are being badgered. Your choice situation is altered, you can either continue the interaction or just take the flyer and be on your way. Taking the flyer is the easiest option. Therefore, you are also being influence by means of manipulation by ideal response.

I do not question that manipulation can take the form of both ideal and non-ideal response. However, I would point out that the former is a lot simpler. Imagine I wanted people to stop people from using my favourite room in the library. To do this, I alter the air-conditioning for the room when I am not in it, making the room quite hot. It is not so hot that it is intolerable, but hot enough that people tend to avoid it for cooler rooms. Whenever I want to use the room, I just adjust the air-conditioning. In changing the air conditioning, I alter the choice situation of people considering using the room. I make it so there is a negative consequence for using the room, occupants having to suffer the cold. The consequence is not so severe that I force people not to use it, but it makes using that room a worse option. This is manipulation as pressure, and manipulation by ideal response. Nobody need be aware that I am influencing them in this way. Nothing about manipulation as pressure necessitates the awareness of the person being manipulated as to the fact they're being influenced. Only that their choice situation is altered in the correct manner.

What this example demonstrates is that manipulation as pressure can be extremely straightforward and uncomplex as far as influence goes. In the example of the room, I could just as well lock the room and achieve the same effect. Somebody could open it, if they were willing to force it open and run afoul of the librarian and university security staff, or they could summon the effort to ask one of the library staff to unlock it for them. I have again intentionally altered their choice situation to stop them using the room. This sort of influence is almost indistinguishable from any other influence, it is only more effective. For this reason, I would suggest we identify these examples as manipulative because we imagine the people who use this influence as interfering with our goals in an improper way. We may imagine scientists who lead a mouse through a maze for an experiment, changing the walls and paths the mouse can take. Nobody in this example wants to be the mouse. We want to make choices freely, without the interference of a sort of overlord, who guides our decisions. While manipulation as pressure is influence worthy of the title, I would distinguish it from manipulation as bypassing or subverting reason. The former is a simpler influence, the latter more engaged with our mental states, our goals, reasons, desires and emotions. Therefore, it is the latter which I will engage with for the remainder of this project, and what I think of as the, for a lack of a better term, 'true' concept of manipulation.

Summary

In summary, manipulation as bypassing or subverting reason is the starting point for many definitions of manipulation. The more interesting questions are what exactly it means to bypass or subvert reason, and how a definition can be crafted which fits the many apparently different examples of influence we intuitively identify as manipulative. The theory I want to advance in the next chapter is what I think has the best chance of answering these questions satisfactorily.

Manipulation is important, and it is an interesting category of influence to explain and analyse. Not least because the ethical character of manipulation is of such interest, and it appears to be so common in different elements of daily life. We also know it is not enough for a definition of manipulation to only explain manipulation on a single scale. A definition should not only allow us to explain small scale interaction between friends for example. It should also allow us to understand exactly how politicians influence millions through a televised speech, and how the systems we interact with through government or technology influence is in way we consider manipulative despite there being no single agent behind them.

Chapter Two: A Definition of Manipulation

In **Ticket Deception,** Mr Fisk is received by the ticket salesperson regarding which platform his train will depart from at the station. Consequently, Mr Fisk boards the wrong train. He travels many miles until he finally realises he is on the wrong train. Imagine that he hears the train conductor announce the next stop, examines his watch and it dawns on him that he should have arrived at his destination by now. Imagine that he frantically stands and asks the train conductor to repeat themselves, and when he does realises he has never even heard of this station before. On confirming he is on the wrong train; Mr Fisk recalls his conversation with the ticket salesperson and how he treated them. He curses. He's been tricked.

Consider another scenario. Imagine the deep pit of unease and humiliation which settles into an elderly gentleman's stomach when the bank teller informs him that his account has been emptied. The old man received a call the previous night from an unknown number. The person on the end of the phone told the man that he had won a large sum of money in a prize draw. All he needed to do to claim the money, they said, was give them his bank details so they could transfer him the money. He was warned that if he didn't do so immediately, the prize draw could rollover and he wouldn't receive it. Elated, the old man gave them whatever information they asked for, including his PIN number and security codes. As he is told again by the bank manager that his account is empty, realises that he has been tricked. The person on the other end of the line was a con-artist. He has been tricked out of his life savings.

In the previous chapter we examined how accounts of manipulation have developed from a key starting point. This starting point is the intuition that to manipulate is to influence a person in a way that is starkly different from persuasion. Whereas we persuade a person by engaging with their capacity for reason, we manipulate them by either bypassing this capacity, or subverting it. The difficulty in most accounts in this school of thought is that 'subverting' is rarely defined to a standard adequate to our needs. However, Robert Noggle's account of manipulation as trickery is one which I argue does define manipulation convincingly as a development of the other accounts. It gives further clarity on exactly what it means to subvert reason and does so in an elegant way which adequately explains our intuitions regarding the wide variety of ways manipulation appears to operate.

That does not mean that the trickery account as presented by Noggle is free of flaws. Yet, I argue these flaws can be overcome. Firstly, however, I should present his account.

Noggle's Definition of Manipulation as Trickery

In his paper '*Manipulative Actions: A Conceptual and Moral Analysis'*, Robert Noggle develops an account of manipulation as trickery which is designed to account for a variety of examples. In the interests of space, I will reiterate here only those which I view as the most useful three. I will also ascribe to them my own monikers for ease of reference:

Shakespearean Anger: "Iago plays upon Othello's anger and jealousy, so that when he becomes convinced that Desdemona has been unfaithful, he becomes enraged and murders her."[10]

Biblical Temptation: "Satan tempts Christ who is fasting in the wilderness. He reminds him of his hunger and of the fact that he could turn the stones into bread."[11]

[10] (Noggle, 1996, p. 43)
[11] (Noggle, 1996, p. 43)

Financial Scam: "Claiming to be a police officer trying to catch a dishonest bank teller, the swindler asks his victim to withdraw her life savings. He then absconds with the money."[12]

Each of these Noggle takes to be a case of manipulation. Furthermore, each operates in a different way to the others. In the first case, it is the emotion of the target agent that the manipulator, Iago in this case, intends to influence. In the second example, Satan intends to influence Christ by inflaming his desire, in the biblical case, his hunger. Finally, in the third, it is the beliefs of the target that the con-artist intends to influence. This forms the basis for the account.

Noggle asks us to imagine that people are driven by three cognitive levers, which the manipulator acts to push and pull as desired. They are our beliefs, desires, and emotions. To make what we regard as 'good' decisions, we seek as a matter of practical reason to organize these three elements according to particular 'ideals'.[13] These can be expressed as propositions we strive to attain. We can refer to agents as being more or less 'ideal' in respect to the extent to which they meet these standards. These are clearly aspirational, hence Noggle's use of the term 'ideal' in the first place. It is not as if people are not perfectly capable of being 'unideal', for example by holding false beliefs, without being influenced to be so. Noggle also calls them "norms"[14], suggesting that they are prescriptive. A rational agent should, or at least we expect them to, strive for these ideals. When an agent is made less ideal, they are "lead astray".[15]

One of the main challenges of defining manipulation has been producing a definition which can accommodate for these sorts of examples where manipulator's utilise different aspects of our persons to influence us. Noggle therefore defines manipulative action as "the attempt to get someone's beliefs, desires, or emotions to violate these norms, to fall short of these ideals"[16] I can state this formally, with some caveats:

Manipulative Action 1 (MA1): An act of intentional influence directed at an agent, where the actor acts with the intention that the agent falls short of one or more ideals which govern their beliefs, desires or emotions.

Let us note some important aspects of this definition. Firstly, it is a definition of manipulative action. It is not meant to be a definition of manipulation only when it is successful. It makes no mention of under what conditions the manipulator succeeds in actually manipulating the agent, nor under what conditions they fail. To act manipulatively in this account is to attempt to manipulate, regardless of success. Though we can perhaps further distinguish between an act of manipulation which is an attempt, and attempting to attempt to manipulatively but failing even to act (perhaps the influencer tries to speak, and so manipulate, but is cut off before they get a chance). For now, it's enough to say this is a definition of what makes a particular act by an agent a manipulative one and leave it at that.

Secondly, I have written MA1 to require that the influence is intentional. Noggle clearly states in his paper that it "seems to be crucial to making an action manipulative: that it is done with a certain kind of insincere, conniving intention."[17] It's therefore clear that Noggle considers accidental, or unintentional manipulation to be an impossibility. We can say more on this, and it is worth

[12] (Noggle, 1996, p. 43)
[13] (Noggle, 1996, p. 44)
[14] (Noggle, 1996, p. 44)
[15] (Noggle, 1996, p. 48)
[16] (Noggle, 1996, p. 44)
[17] (Noggle, 1996, p. 48)

considering exactly what it means to possess the intent required for an action to count as manipulative. I have included it in this first definition, but this could be clarified.

Thirdly, we should also note that Noggle does not comment on whether or not the manipulator must also have a further goal, or end in mind which they want to achieve as a result of the manipulation. Therefore, I have not included this in MA1. Though we could take this to be implied on some readings of Noggle. To briefly consider the point, I would argue that acting in the service of an external end, an end further to manipulating for the sake of it, may not be an essential element of the manipulation as trickery account. Consider a prisoner who is serving a long sentence in custody. Perhaps they try to manipulate their jailors by lying about events in the prison, such as the location of violent incidents. They do not do this for an external aim, in the way we could imagine a prisoner trying to manipulate a guard to smuggle them contraband or some such other goal. Rather, the prisoner simply finds enjoyment in manipulating those who have power over him. This seems to be a close as one can come to manipulating for the sake of it, rather than an external goal. Others may disagree and argue that deriving a sense of power from manipulation is an external goal itself. Regardless of the conclusion we come to, it is tangential to the definition, rather than essential for practical purposes. Either manipulation for manipulation's sake is possible, in which case having an external goal in mind is not necessary, or it is not, in which case it is. I would argue that manipulation can be done for its own sake, so much as agents can have no other goal in mind other than making a person less ideal, but it is not essential to the definition.

With these immediate thoughts in order, I will now critically examine and explain Noggle's definition in greater detail, before moving to my broader criticisms of it, and then my proposed developments to the definition.

Critical Analysis of the Manipulation as Trickery Account

Noggle gives detailed explanations of what the ideals are for our beliefs, desires and emotions, and how manipulators can attempt to influence an agent with the intention that the agent falls short of the ideal. Let us examine them in turn.

Noggle claims that two ideals govern our beliefs, though they can also be referred to as a single, conjunctive ideal if one so chooses. I believe it is easier to understand and apply it if we refer to them as two separate ideals. The first is that agents should only believe that which is true. The extent to which an agent holds false beliefs is the extent to which they fall short of this ideal.[18] Consequently, to intentionally lead a person to form a false belief is to make them less ideal in this regard, and thus manipulative. Obvious cases of this sort of manipulation are those where an agent directly lies, or tells a series of lies to another person, as in **Financial Scam** and **Ticket Deception**. However, deception can obviously be indirect. For example, we can use implication, leading questions etc. to draw a person to make assumptions that are false without ever directly asserting a falsehood.[19] It is of note that Noggle asserts that deception should be subsumed into manipulation. To intentionally deceive is, according to the trickery definition, to manipulate. At first this may sound radical, but others such as Shlomo Cohen make convincing arguments for deception to be considered a subset of manipulation.[20] Deception, in so far as we define deception as influence which is intended to make another agent fall short of the ideal for belief, is then a manipulative act.

[18] (Noggle, 1996, p. 44)
[19] (Noggle, 1996, p. 44)
[20] (Cohen, 2018)

The second ideal is that agents should only attend to those beliefs which are relevant to the context at hand.[21] When trying to utilise one's rational capacities, an agent will only utilise a select set of beliefs which are relevant to a given context. If I am trying to decide what to eat, I would not usually hold in my mind the belief I have as to how many planets there are in the solar system, for example. That belief, true or not, is not relevant to the given context. This filtering of our beliefs is something we perform for the obvious reason that our capacity to pay attention to beliefs we hold is limited. We can be influenced to attribute more or less importance to a belief than the manipulator believes that belief is due, and thus pay it less or more attention. We can be flooded with irrelevant information such that important information is obscured, or otherwise made to pay attention to beliefs which may well be true but are misleading. An agent falls short of this ideal if they do not attend only to relevant beliefs in a given context. Conjunctively, we could say that one should only attend to both relevant and true beliefs in a given context.[22] However, if I influence a person to believe something I think is false, they have been "led astray"[23] regardless of the situation or context we occupy. Therefore, I believe it is best to separate the two.

Moving onwards to consider desire, Noggle assumes that human motivation is "instrumentally rational."[24] By this he means that our desires "usually conform to our beliefs about what we have reason to do."[25] This should not then be confused with how the term is used in other philosophical contexts. Moreover, this is the ideal. We want our desires about what we want to do, to accord with what we have reason to do. Of course, sometimes they do not. We desire things we know, on reflection, we have no reason to desire, or actually reason we certainly should not desire. We're often perfectly capable of holding desires we do not have reason to hold without any outside intervention from hostile agents. However, we can also be influenced in what Noggle sees as two distinct ways which can cause us to fall short of the ideal. The first way is through conditioning.[26] A person can be conditioned to form a desire with no rational basis at all. A typical example is of Pavlovian conditioning. Scientists ring a bell whenever they feed a dog. Soon, the dog begins to salivate whenever a bell is rung, regardless of whether the dog is given food or not. Conditioning can also create desires which are aligned with a corresponding belief. For example, a person with anxiety may be conditioned by their psychologist using repetitive mantras to alleviate feelings of inadequacy in social situations. In this case, conditioning does create a desire to motivate what the agent has reason to want. However, conditioning obviously can produce desires with no rational basis. Doing so is making the target of the influence less ideal and is therefore manipulative. For example, a political prisoner may be beaten when they speak their native language, and soon this fear is internalized, even when the threat of beating is removed, and they are released back to society as a 'reformed' person. They no longer have a rational reason which conforms with the desire to not speak their native tongue. There is no guard within hearing distance. Yet they have been conditioned to have the desire to only speak the desired language of the regime. The beatings are examples of coercive influence, but also manipulative influence in this case. There is also the example in my first chapter of **Conditioning by the Party**. Through constant repetition of nationalist slogans, Smith forms many beliefs, but also an internal desire to act in nationalistic ways which are not connected to his beliefs.

[21] (Noggle, 1996, p. 44)
[22] (Noggle, 1996, p. 45)
[23] (Noggle, 1996, p. 44)
[24] (Noggle, 1996, p. 45)
[25] (Noggle, 1996, p. 45)
[26] (Noggle, 1996, p. 45)

The second application of this ideal involves competing desires. We do not just want to possess a desire connected to a belief about what we should do. We also want that desire to have sufficient motivational force to move us to act as we have reason to do. Desires can be too weak, especially in the face of competing desires. An example would be if you had a belief that you should go to the gym each morning, as you value your health. However, the desire to do this may be weak, and the desire to stay in bed for longer stronger. Noggle identifies this as "motivational akrasia".[27] He defines it as when "one lacks sufficient motivation to do what on believes there is reason to do".[28] Either the person lacks any desire at all to do what they have reason to do, or their strongest desire is not for what they have most reason to do, compared to competing desires and any associated reasons they have. This state accurately describes the consequences of successful temptation. Temptation is a good example of how a manipulator might inflame a desire which motivates an agent in a way which the agent does not have reason to do. In **Biblical Temptation**, Satan tempts Jesus to use his divine powers to feed himself, which is not what Jesus has most reason to do (continuing his fast in the desert). In my own example, **Peer Pressure**, I believe Percy tempts Sarah. Sarah has a reason and an associated desire to abstain from drinking. Peter tempts her by buying her a drink and placing it in front of her. Sarah's desire to drink, which she has the strongest reasons not to act on, is inflamed by this action. We can also understand as manipulative influence which intends to make a desire less potent, when the manipulator believes the target of the agent has the most reason to act on that desire. Discouraging desires can be just as manipulative as inflaming them.

Finally, there are also two ideals for emotion. The first is that we should ideally only feel emotions which are appropriate to the context. The second ideal is that we should feel only emotions which make more salient features of the world which warrant such salience. Let's consider appropriateness first. I would argue that, to an extent, the ideals for emotion mirror the structure of the ideals for belief. An emotion can be considered appropriate if it is fitting that an agent should experience it within a given context. By the given context, I mean the situation which has caused the emotion to be felt. Noggle's description of when an emotion is appropriate is that "in general, "positive" emotions such as joy, happiness, hope, and so on, are appropriate for someone who desires that *P* and believes that *P* is true of likely. Negative emotions such as anger, fear, and so forth, are appropriate for someone who desires that *P* and believes that *P* is false or unlikely."[29] The conditions which are *P* will differ for each emotion, and we do not have a complete picture as to what these conditions are, and whether or not this is an objective set of conditions or a subjective one. Nevertheless, it is ideal to only feel emotions which are appropriate. Imagine if I am told that one of my friends has betrayed my trust. I could feel anger directed towards my friend, because I believe they have betrayed my trust. It is appropriate for me to feel this way if it is actually true that my friend betrayed me. If on the other hand I find out after speaking to my friend that he did not actually betray me, and I was lied to, it would not be appropriate for me to continue feeling anger towards him. Moreover, I have been manipulated by the one who told me the lie. It was actually never appropriate for me to feel anger at my friend, and thus I was made to fall short of the ideal by the manipulator.

I have three things to note about this ideal. The first is that Noggle's appropriateness is essentially the concept of fittingness, an example of which is "fear is rational in terms of fittingness just in case it is directed towards things that are truly dangerous".[30] Crucially, that an emotion is appropriate in

[27] (Noggle, 1996, p. 45)
[28] (Noggle, 1996, p. 45)
[29] (Noggle, 1996, p. 46)
[30] (Scarantino & de Sousa, 2021)

the sense that it is rationally fitting, does not mean that it is ethically justified, or justified relative to other standards. The emotion can be justified in the sense of being appropriate given the context and facts, while unjustified from an ethical perspective. A common example is a sexist joke. It may be fitting to find a joke amusing, but unjustified ethically.[31] The second element to note is that this ideal is essentially the same as the ideal of truth for belief. Appropriateness may have a great deal more subjectivity than truth, at least on the face of it. However, practically speaking they are quite similar. Appropriateness ultimately is the eliciting of an emotion on the basis of a true or false belief the manipulator convinces their target to believe. I should also note that this is a cognitive approach to emotions, which was referred to in the first chapter. **Shakespearean Anger** is the example Noggle uses to demonstrate this form of manipulation, although it also has many elements of the others. Iago influences Othello to feel an emotion which is inappropriate by its scale, and thus Iago manipulates Othello. This ideal of appropriateness runs parallel to the ideal of truth for belief. Our emotions should be rational responses to reality.

Finally, the second ideal for emotion is salience. Noggle draws on the work of Ronald DeSoura, another philosopher who has explored the rationality of emotions. He argues, following DeSoura, that emotions function to draw our attention to important features of our experience. He states that "Joy makes us pay attention to the fact that we've just matched all six lottery numbers. Fear makes us pay attention to the tornado coming our way. Love makes salient the fact that the person who just walked into the room is a loved one."[32] This, I should note, is like the ideal of relevance for belief. Emotions then, ideally, should bring to our attention what are important features of our experience, it should make them more salient. To cause a person to feel emotions which bring into focus irrelevant facts and features about the world is to cause them to fall short of this ideal and thus to manipulate them. The example Noggle uses to illustrate this is that of two friends (Hansel and Gretel) who are discussing where to go to eat. They have an agreement that each will take turns to decide at which restaurant they should eat. When it is Gretel's turn to decide, but Hansel does not like Gretel's choice, Hansel decides to sulk. As Noggle states: "Hansel thus may not set out to change the *content* of any of Gretel's beliefs, for he may know that Gretel is aware of his preferences. Rather his goal is to play on Gretel's sympathy in order to get her to assign greater *relevance* to his preference than it warrants in a time and context in which it is her preferences that should be most relevant."[33]

We have then, a definition of manipulative action, and an explanation of what it means to make a person fall short of one or more of the ideals for belief, desire or emotion. The trickery account defines 'subversion of rationality' as causing a misstep in an agent's practical reasoning. This can take place by targeting the agent's beliefs, desires, or emotions to make them less ideal for practical reasoning. Now, there are some key elements of the definition which need immediate examination.

A concern which immediately springs to mind, especially when considered the ideals of appropriateness for emotion and truth for beliefs, is how we are to decide what is or is not a fulfilment of the ideals when it is possible to reasonably disagree.[34] It is easiest to demonstrate this if we limit ourselves to the ideal of true belief, and apply it to an example like religious beliefs. You and I may reasonably disagree on the question of whether the Christian God exists. To the atheist, Christianity and all religion may seem like a manipulative exercise, tricking people into believing that following religious prescriptions will lead them to an eternal life. On the other hand, a devout

[31] (Scarantino & de Sousa, 2021)
[32] (Noggle, 1996, p. 46)
[33] (Noggle, 1996, p. 47)
[34] (Noggle, 1996, p. 47)

Christian may view the atheist as a tempter and deceiver, who manipulates vulnerable members of the flock into losing their faith, thus condemning them to eternal damnation. Belief in the Christian God, or disbelief, are both reasonable perspectives to hold. Yet both cannot be true. Regarding emotion, two reasonable agents could disagree on whether an emotion is appropriate in a given context.

Noggle answers this question by considering three possible approaches for how we evaluate what is or is not ideal relative to these standards. We can say that the content of ideals is objective, that they are relative to the manipulator, or relative to the target of the influence.[35]

We can first consider relativizing the ideals to the beliefs of the manipulated. On this approach *a* acts manipulatively towards *b* if *a* acts in a way which makes *b*'s beliefs, desires or emotions less ideal, relative to what *b* considers to be the ideal state for their beliefs, desires and emotions. If we take this position, if *a* attempts to get *b* to believe something which *b* considers false, then *a* is acting manipulatively. This is regardless of whether *a* sincerely believes the propositions he expresses. This makes manipulation incompatible with accounts of deception.[36] If *a* intentionally lies to *b* by convincing them of *P*, but *b* already believes *P* and views it as true, then from *b*'s perspective, they have not been made less ideal, while *a* did act deceptively. If it is this perspective, we are using to decide whether an act is manipulative, then one can directly deceive another person and, if unbeknownst to the deceiver their target is already deceived, no manipulation takes place. I believe we have an interest in maintaining the existing link to lying, that one manipulates when they directly lie. Otherwise, we would not count as manipulative a great many examples we could consider intuitively manipulative should we amend the example. While we could accept that if a person does not change their beliefs, a manipulator is not successful in manipulating, it seems right to say they still act manipulatively if they believe they are actively, directly deceiving their target.

We can then consider the objective approach. Noggle does not expand much on this, simply stating that in regard to appealing to objective standards for matters like the appropriateness of emotions we can "hope that the problem of formulating them does not turn out to be intractable, and then either attempt to construct the appropriate ideal or wait for someone else to do so. This strategy does not seem promising."[37] We can expand on this, providing more reasons to reject this approach. If we were to assume that the content of the ideals are objective, then presumably they should not change. At least we might say this should apply to the first ideal for belief (that all our beliefs should be true). Historically, scholars thought that human health required a balance of the four humors, blood, phlegm, yellow bile, and black bile. This is false, yet the scholars knew no better, and we may presume they had good reason to believe as they did. If we take the objective approach, then by our definition, those masters teaching their apprentices in the theory of the humors were acting manipulatively, despite their honest intentions. They were always expressing falsehoods and convincing others of those falsehoods. We can also return to the example of religion. To a Christian, the existence of the Christian God is true, and they may believe they have an obligation to convince others of this truth. Of course, many different religions have contradicting beliefs about the existence, or lack thereof, of any number of deities. Therefore, it follows that whatever the objective facts are in regard to religious questions, at least some sincere, religious folk must believe falsehoods. Therefore, at least one must be acting manipulative when they attempt to convince others of their perspective, should we adopt the objective approach. We would not usually describe manipulators as acting honestly to the extent which calling all religious activity manipulative in the

[35] (Noggle, 1996, p. 47)
[36] (Noggle, 1996, p. 47)
[37] (Noggle, 1996, p. 47)

same we would call say, an abusive husband manipulative, is to me hyperbolic at best and absurd at worst.

We are left with relativizing the standards of the ideals to the perspective of the manipulator themselves about what is appropriate for the person being influenced. This is the perspective Noggle also ultimately takes, though I have developed further than the reasons he gives.[38] The manipulator has beliefs about what is true, relevant, appropriate etc. When influencing others with the intention of getting them to fail in regards to one of the cognitive norms, they then apply their own understanding of what is true, relevant, appropriate etc. to that person. What are important are the beliefs of the manipulator about the target they intend to manipulate. For some ideals, like truth, this would apply universally from the perspective of the manipulator for any person they wanted to influence, as well as themselves. What a person believes is true they may take to be true for everybody. For conditioning and temptation, the manipulator has to rely on their knowledge of the desires of their target. This also preserves the notion that a manipulator may have double standards. They might think anger an appropriate emotion to feel in a context they themselves occupy, but not appropriate for another person to feel in the same context. This also preserves the notion that a manipulator may have double standards.

At this stage, we can therefore add the definition of manipulation:

MA2: An act of intentional influence directed at an agent, where the actor acts with the intention that the agent falls short of one or more ideals which govern their beliefs, desires or emotions, ***relative to what the manipulator believes are ideal for the target agent****.*

This is as far as Noggle goes in defining manipulation as trickery, moving on to consider this ethics of manipulation as trickery. We will do the same in the next chapter, but I believe we can push this definition further. I will start by saying there are good reasons to think that manipulation must be intentional. That is, as we have just discussed, *MA2* states that in order to manipulate, the influencer must have an intention of the correct form towards the target of their influence. This can be useful, albeit it forms the basis of two criticisms. One I will consider in the next part of this chapter when considering challenges to the definition and how they can be resolved without amending the definition. However, there is one which is worth tackling now, to assist in developing the trickery account further.

Manipulation as Trickery and Intention

The immediate criticism of this aspect of the definition then, is that we can question whether apparently manipulative agents have such an intention of the right form to count as a manipulative intention. Noggle presents the "intentional" action of the manipulator as a sort of planned intent. The manipulator first forms a plan of how they will influence a person, then puts that plan into action. The problem is that many manipulators may not be making such plans, but still engage in behaviour we want to say is manipulative. Additionally, we can imagine cases where at the time, an agent does not think they are being manipulative, but later comes to think of their actions as manipulative.

Manipulators are unlikely to plan their actions in the same philosophical terms we have been using. A manipulator does not think in terms of making their target less ideal. For truth and relevance, they may think of lying, misleading and distracting. For the emotional ideals, they may think of themselves as simply hurting another person or drawing attention to themselves. Instead of

[38] (Noggle, 1996, p. 47)

conditioning, they may think of themselves as only 'training'. Temptation may keep the same terms, as it's a familiar and non-technical concept. Of course, the term does not matter as much as intention to perform an action we'd identify as being manipulative, or intention to act in a way which clearly will result, if successful, in the target falling short of an ideal, to the extent where the agent would definitely know that will be the outcome.

Noggle uses an example of Hansel and Gretel to demonstrate when a person makes another fall short in regard to the ideal of salience for emotion. Hansel and Gretel are two friends who go out each Friday. The two friends take turns each week deciding where they should eat. Gretel chooses a restaurant that Hansel is not fond of. Hansel huffs, sulks and complains that the restaurant is terrible, and Hansel will have a terrible time there. Seeing how his mood has turned, Gretel sighs and asks where he wants to eat. Gretel is manipulated here by Hansel because Hansel elicits Gretel's sympathy, with the view of making his opinions on where they eat more important than the context should allow. They had an agreement to take turns, and thus value each person's preferences on where they eat equally, but Hansel tips the scales in his favour by making Gretel feel an emotion which makes more salient Hansel's feelings about where they should eat over Gretel's feelings, as well as their any obligations arising from their arrangement.[39]

We could think that Hansel had a sort of planned intent here to sulk in order to renege on his agreement. But perhaps he is not that self-aware, and this is just how he acts when he does not get his way. Suppose now that another friend has joined them in their excursion, Agatha. When Hansel sulks, Agatha steps in and accuses him of acting manipulatively. Hansel disagrees, saying he was just reacting to the prospect of a meal he wouldn't enjoy. He wasn't trying to change where they were going. Agatha explains to him that despite his and Gretel's agreement, he was unfairly playing off of Gretel's sympathy for him to make her put his preferences above his own. After this discussion, Hansel begrudgingly agrees it was unfair of him, and he shouldn't be so expressive and sulk when he doesn't get his own way. Rather, he should try and enjoy the meal. After all, he's probably taken Gretel to many places she didn't enjoy as much as he did. It's only fair that he tries and make the best of it. He now recognises that he was acting manipulatively, and apologies to Gretel.

Hansel has to admit that he was intentionally influencing Gretel, but he does not realise this until it is explained to him. The problem this example presents for *MA2* is that we may run into difficulty attempting to claim that Hansel had manipulative intent, when it seems he was entirely unaware of the effect of his actions, let alone how they operated to realise that effect.

What this example shows is that manipulators do not need to think they are manipulating. We must accept that. How then, are we to develop *MA2* to capture this fact?

We can first note that while manipulators may not necessarily think they are acting manipulatively, they must be attempting to influence intentionally. To intentionally attempt to influence necessarily entails they must be aware that they are engaging in influence. They do not need to be precisely aware of how that influence functions. We should also consider that Gretel can examine his own intent retrospectively, perhaps changing his evaluation of his past actions. Consider another example which better highlights this aspect of the relationship between intent and manipulation.

Reformed Abuser: Joe and Karen are married. Joe consistently makes Karen feel inappropriately negative about herself. She tells him she's being considered for a promotion at work, and he tells her that it's probably because her supervisor finds her attractive, not because she's actually good enough for the position. She wants to go out with her friends, but he tells her they talk about her

[39] (Noggle, 1996, p. 46)

behind her back so she shouldn't go. She tries on new clothes, and he tells her she looks disgusting in them, and should stick to her existing clothes. This continues until Karen eventually seeks help from friends who identify that Joe is emotionally manipulating her. She consequently initiates divorce proceedings against him. Joe vehemently denies he was acting manipulatively and begins to stalk her obsessively. He is arrested and found guilty of stalking and as part of his sentence, he is forced to engage in a programme aimed at reforming those engaged in domestic abuse. As part of this programme, he receives counselling and realises that the reason he constantly attempted to lower Karen's self-esteem was because he had low self-esteem himself and was obsessed with making himself feel stronger. He now realises he acted manipulatively.

Joe acts without a planning intent of which he is consciously aware. Rather he acts with intent he is not aware of until he reflects. There is no doubt that he acted with some sort of intention. He did not think at the time that he was not intentionally influencing Karen, in the same way a person might unintentionally act if they reacted instinctively to an object speedily approaching their face. The question is whether it is acceptable to say that he acted manipulatively, without knowing that his intention was one of manipulating.

We're often asked to consider why we did an act *P* way, where *P* is some past action we're asked to justify or reflect on. It is through this process of reflection that we realise facts about ourselves and those past situations that we were not conscious of at the time. If we lash out angrily on one occasion, we may later reflect on why we lashed out. When we identify *why* we performed an action in the past, this previous motivation can be called a backwards-looking motive. Providing a full picture of what it means to have intention, or what it means for an act to be intentional, is an extremely complex topic that does not fit within the scope of this thesis. The important question is what sort of intent would be required for a person to act manipulatively.

We could conceive of Joe as, contrary to the example, never viewing his intentions as manipulative at any point in time. Can we then still say that he had the sort of intent which would allow us to categories his acts as manipulative, as we may believe any competent account of manipulation should, even if he never reflects on his actions?

I believe we can. I argue that to in order to categorise the sorts of actions in examples like **Reformed Abuser** using the trickery account, we must assume that while a person cannot manipulate unintentionally, they can be unconscious of the depth of their intention at the time they act. Rational agents are clearly able to reflect on their past mental states and attitudes. Consequently, they are also able to identify past reasons for their actions that they may not have been aware of in the moment before acting, which nevertheless formed a part of their motivations and intent which caused them to act as they did. Consider, through therapy a person might find a whole host of reasons for their past behaviours they were not aware of, or that they did not fully understand, until reflection with a therapist or counsellor. This is a different result than a person whose attitudes have changed, such that they now evaluate the ethics of their previous actions differently. For example, a person might believe *x*, and convince people of *x* but then later in life no longer believe in *x*. While they can coherently express regret at their beliefs and how they spread that false belief, they cannot retrospectively decide that they were lying, unless they are asserting that they were, to some extent, aware of the falsehood of *x* at the time.

We can safely say then, that it is not a necessary condition of manipulative acts that the influencer must be aware at the relevant times of intending and acting that their act of influence against an agent *A* will cause *A* to fall short of an ideal from the perspective of the influencer. Rather, it is necessary only that the influencer would, at the relevant times, believe that their act would cause *A*

to fall short of an ideal and nevertheless acted intentionally to influence them. In order for Joe to accurately identify himself as acting manipulatively when evaluating his past actions, these three conditions should hold:

C1) Joe must accurately believe that he held attitudes regarding what was ideal for Karen's beliefs, desires and emotions at a time *X*.

C2) Joe must hold that his influence at a time *X* was aimed at causing Karen to fall short of one or more of those ideals according to Joe's attitudes about them.

C3) Joe must hold that he intentionally influenced Karen.

If these hold, then Joe can evaluate himself as being manipulative in the past, even if he was not aware at the time of the manipulative reasons which drove his actions. Put shortly, he can accurately believe that he emotionally manipulated Karen even though he was not aware he was doing so at the time.

This train of thought does not entail need to change *MA2* further without making additional arguments. Rather, it clears obstacles for a final change to the definition based upon further reflection on the link between manipulation as trickery and existing theory about deception. To summarise, if we consider examples such as **Reformed Abuser** and Noggle's own example of Hansel and Gretel, we can easily consider cases where it appears somebody committing what we may regard as clearly manipulative acts does not have the sort of intent present in manipulators in other examples. The intent in those latter examples is planned while in the former the intent is less detailed and only on reflection would those agents identify the true scope of their intent. From this we can conclude that, without providing a full account of what it means for an act to be intentional, that agents do not necessarily need to be aware that their intent is manipulative, in order for their intent to be manipulative. More precisely, they do not need to believe, at the relevant times before acting to influence another agent, that they will be acting manipulatively, nor even that they will intentionally make the agent fall short of an ideal. Rather, it only needs to be the case that as a result of their beliefs at the time about their own actions and the target of their influence, that they *should* believe it as a logical consequence.

Manipulation and Bullshit

I argue that manipulators do not need to know they are acting manipulatively in order to act manipulatively. To repeat, this does not mean a person can manipulate unintentionally, or accidently. Rather, a person can have imperfect knowledge of their own intentions, yet still act with that intention.

As already stated, the way that manipulation as trickery is defined entails that all intentional deception is also manipulation. As a consequence, it may follow that we can examine other forms of deception, along with the theories that explain them, and consider how an account of manipulation may include them. One of these forms is expressed by Harry Frankfurt in his influential paper '*On Bullshit'.* Frankfurt characterises bullshit as distinct from lying. While a liar knows what they express is false when they lie, the bullshitter has either no knowledge or no regard for whether what they express is accurate. The liar is extremely familiar with the truth value of the statements they lie about. The bullshitter has no regard for the truth of their statements at all.[40] Similarly we might think of manipulators whose influence is planned and targeted to make an agent fall short of an ideal, and a manipulator who has no regard whatsoever for the affect they have on the target beyond

[40] (Frankfurt, 2005, p. 19)

achieving the practical goal of the influence. I would estimate that the latter is a much more realistic picture of what we might say is in the mind of most manipulators. The manipulator won't really be considering the truth of their own statements, or whether an emotion they are trying to elicit in their target is inappropriate for their target. Rather, they're thinking about the goal they want to achieve through their influence, and influencing with whatever tactic seems to work, regardless of any other consequences. This sort of manipulator isn't well considered by MA2. Those sorts of manipulators tend to be careful; this sort of manipulator is reckless.

If we adopt the possibility of reckless manipulation, then manipulative influence is slightly different from MA2. Consider the following example:

Vegetable Soup: John works as a waiter in a restaurant. One of his customers asks if the vegetable soup on the menu is vegan. The restaurant is very busy, and John is eager to take this table's order and move on to the next one. John does not know whether the soup is vegan but tells the customer the soup is suitable for vegans. After all, its vegetable soup, without any meat. Later, he realises that the soup contains butter, which makes the soup not suitable for vegans.

Did John manipulate the customer? In one way he certainly does, as while he tells the customer the soup is vegan, by doing so he also implies that he knows the previous statement is true. If John had said he did not know, he would have avoided making the customer less ideal. More importantly the answer itself, that the soup was suitable for vegans, was manipulative. Yet we cannot call it a lie. John did not know whether the statement was true or not, nor did he care. At the moment of his utterance, he only cared about answering the question quickly in a way which would allay the customer's concerns and allow John to move on to the next table. Yet we cannot say that John intentionally led the woman astray as would be needed for this to count as manipulative under MA2. For all John knew, he was telling the truth.

It appears then that we need to amend the definition. I would propose that the correct way to do it, is to turn the definition on its head. It is not just that in order to manipulate, an agent needs to intentionally make another fall short of an ideal. Rather, agents who intentionally influence other agents, *without* trying to *avoid* doing so, act manipulatively.

When an agent is not acting manipulatively, but still acting intentionally to influence another agent, we should expect that if they are attempting to influence us without manipulating us, they will have some regard for how their influence affects us. We could perhaps think of it like a martial arts spar or contest. There is an understanding that while individuals are trying to affect the other, in the martial arts contest physically, while in a context of influence mentally, there is an element of care for each-others well-being. A person who tries to influence me in good faith, does so with at least some minimum effort to avoid making me less ideal. The manipulator does not, either because they are being reckless and do not care what state I am left in, so long as they achieve the results they desire, or alternatively because they are to an extent conniving, and purposefully lead me astray.

Manipulation then, is still trickery. It is just that leadings others astray need not be so active a pursuit. An agent can just as easily trick due to negligence or recklessness in how they act, just as well as they can if it was wholeheartedly purposeful. We should therefore amend the definition of manipulative action as follows:

MA3: *An act of intentional influence directed at an agent, where the actor acts **without** the intention that the agent **does not** fall short of one or more ideals which govern their beliefs, desires or emotions, relative to what the manipulator believes are ideal for the target agent.*

This is a departure from Noggle's definition, but I believe it is a relatively natural development of the trickery account. Without it, we cannot I think honestly say that the trickery account captures the different types of intent, some detailed, so not very much, that manipulative actors have. Manipulative action can be done without really thinking. Lies told quickly to save one's ego, emotional outbursts where we hurt people without really thinking about it, temptations hurled at people out of a sense of moral superiority, or moral insecurity. All intentional, but only on reflection are these intentions plainly visible. We manipulated yes, but we did so out of a lack of care for how we influenced, so long as we got our way, rather than out of a sort of planned malice.

One thing to note before we move to the various problems and replies with the trickery account is that it is quite unwieldy, both when considering objections originally put to Noggle's trickery account, as well as due to the negative phrasing of MA3, necessary though I believe it is, to apply the definition in full every single time it is applied. Henceforth, when the shorthand of "intentionally making another fall short of an ideal" is applied, it should be noted that properly this also includes cases of reckless manipulation. After all, it is still intentional manipulation, though the intent differs slightly.

Problems and Replies

Since its publication, Noggle's trickery account has faced some counterexamples and criticisms which are also applicable to my altered definition. In the following section I will address what I view to be the most severe of these issues. As a result of adapting to these criticisms, we will gain a more precise understanding of the definition of manipulation as trickery.

The Sincere Manipulator Problem

We can use the example of **Apocalyptic Preacher** to demonstrate this criticism. In that example, there is a television minister who spreads fear of an imminent apocalypse, while also selling supplies which she claims will help her followers survive that apocalypse. An impartial observer might identify the preacher as acting manipulatively. The preacher spreads fear, then profits from the fear they spread. We view them as a charlatan who is knowingly causing their followers to become less ideal, at least by lying to them, if not by other means as well. This assessment, however, assumes that the preacher doesn't actually believe what they are preaching. What if they are sincere?

If the preacher does really believe that the apocalypse is almost upon us, and that their followers will need the survival supplies they are selling, then they do not have the intent required for their acts to count as manipulative under *MA3*. They aren't trying to make their followers fall short of an ideal, rather they actively believe they are making their followers more ideal by convincing them to believe something true.

A similar example is given by Jason Hanna in his paper '*Libertarian Paternalism, Manipulation and Preferences'*.

"'*Sincere Cult Leader:* a religious cult leader recruits followers through the sorts of apparently manipulative (but not quite coercive) tactics often employed by such groups. The cult leader sincerely believes his tactics improve the recruits' capacity to see the truth and make prudent decisions about whether to join the group.'

The cult leader directs his recruits toward the deliberative standards that he views as ideal. But this fact alone seems insufficient to acquit him of the charge of manipulation."[41]

[41] (Hanna, 2015, p. 633)

These examples are used to demonstrate a problem with the trickery account, whether it be Noggle's original version of it or my amended one. As the definition requires manipulator's have a specific intent, agents can act in ways which outwardly appear identical to manipulative action, but with a sincere intent. Their actions fall short of the definition, while potentially having the same effects on the agent's being influenced, and the same harms. The critics conclude then that the definition is underinclusive. A cult leader, employing all the apparently manipulative techniques cults tend to utilise, may have sincere intent, but surely if any set of actions should be defined as manipulative, it is theirs.

Hanna specifically argues the solution is to move the trickery account away from a manipulator-intent definition and instead take the objective standards approach.[42] However, I do not believe this is the only solution available.

Hanna is vague when he refers to "tactics often employed by such groups". Clearly, it is the use of these tactics which are the source of the accusation the cult leader is acting manipulatively. Were they to simply politely approach people in the street and engage them in a dialogue about whatever theology they have adopted, we would see the individual as a kooky religious leader, but not a manipulative cult leader. Therefore, let's take one technique cults may use which are not coercive, but not entirely innocent either.

'Love bombing' is a term originally used internally by the Unification Church, a new Christian religious movement which has been accused of being a cult.[43] to bombard an agent with a greater level of positive attention than otherwise would be expected in the context. This could take the form of encouragement, affection, and other forms of positive attention. Between those not in an existing relationship, the purpose of 'love bombing' is to quickly escalate the degree to which individuals trust and feel affection for each other.[44] These bonds of trust can then be used to influence the target to commit themselves in some way to the cult, at which point the love bombing can cease. Abusive partners may also use this. At the start of a relationship, they may be a perfect partner, ever attentive to their partner's needs. Then, when their target is sufficiently invested in the relationship, withdraw this affection. As part of the cycle of abuse, the love bombing may restart as a form of reconciliation, as a way of convincing the target to stay in the relationship and enforcing their control of the target.[45]

Let's assume our manipulative cult-leader employs this tactic. Assume as well that as Hanna states, the cult-leader believes this will "improve the recruits' capacity to see the truth and make prudent decisions about whether to join the group."[46] We could accept that the cult leader believes recruits will be in a better position to see the truth, and become more ideal in this respect as a result of the love bombing. Perhaps they simply view this as their followers showing potential recruits what loving people members of the cult are. However, love bombing is, by any sensible definition, an exaggerated, insincere expression of affection. Use of the tactic as a recruiting tool is by definition deceptive, as cult-members are being directed to overexaggerate their affection for the purposes of recruitment. This may involve explicit lies about how much they care for and adore the target. We might also think that love bombing causes targets to fall short of the ideal for emotion as well. Love bombing causes targets to form deeper emotional bonds to those influencing them than might be

[42] (Hanna, 2015, p. 634)
[43] (The University of Virginia Library, 2001)
[44] (British Broadcasting Corporation, 2000)
[45] (Archer, 2017)
[46] (Hanna, 2015, p. 633)

warranted considering the short time they have spent together. However, we interpret it, love bombing must be manipulative. After all, if the excessive affection was genuine, it likely wouldn't stop, and thus wouldn't be deceptive at all. In this case, it wouldn't be love bombing.

How does this change our evaluation of the sincere cult leader? Well, while we recognise they may sincerely believe that the use of such techniques makes them more ideal in one respect, sincerely using the technique, with an understanding of what it is, means that the cult-leader or cult-members do intentionally make their fall sort of an ideal, whether by a conniving intentional manipulation or a reckless intention.

I hold that the techniques of a cult are what make the cult manipulative, not the beliefs they are espousing if they are sincerely held. If a cult does not use any of these techniques, then they are not really a cult in the pejorative use of the term. A man who thinks he is the messiah who will save the world may indeed spread this belief using persuasion alone, and no manipulative techniques. In this case, he is the leader of a small, strange, religion. Cults use manipulative techniques because they are effective at spreading beliefs which could not be effectively spread through persuasion alone, or alternatively because the cult leader is not a sincere influencer.

Another example of a technique we may say is unavoidably manipulative, despite sincere intention, is conditioning. To condition is to influence a person to possess a desire without a corresponding relevant belief. A cult which employs regular conditioning against its members cannot claim to not be acting manipulatively, even if they believe the conditioning leads cult members to have beneficial or objectively desirable desires. The cult leader may sincerely believe that cult members who have a desire to be obedient are more ideal yet if they are being objective must acknowledge they were manipulated and made less ideal in regard to their desires, even if they are made more 'ideal' in some respect from the cult leader's perspective. All that is required to make an act of influence manipulative is the intent that the person being influenced fall short of one or more of the ideals, regardless of whether they are made *more* ideal in another respect.

In conclusion, Hanna's sincere manipulator problem can be solved. We don't need to adapt the trickery account so that it defines manipulative action by objective standards. Rather, while it remains defined by the perspectives of the manipulator as to what is ideal for their target, we understand that some techniques, especially those which crop up in sincere manipulator' examples, are unavoidably manipulative if intentionally used on target agents.

Objectivity and Identifying Manipulation

Amending the trickery account to define manipulation in an objective manner, rather than relativising it to the intention of the influencer is something of a reoccurring thought. It will be useful here to consider another smaller problem with the trickery account as I have defined it, which an objective approach would claim to solve.

By relativising the definition of manipulative action to the would-be manipulator's perspective and intention, we may worry that we can never conclusively identify people as acting manipulative, or at least it may be very difficult. An objective approach would not have this problem. We would, as observers to other's behaviour, be able to consider people like the sincere cult leader, or the preacher from **Apocalyptic Preacher.** We could both identify their actions as manipulative as well as identify the consequences of their influence as a result of manipulation without needing to consider the agent's intent at all.

We do not have access to the internal mental states of others. We would never be able to know, with absolute confidence in our belief, whether a person's mental state is one which means they had

intent to manipulate. The criticism is that the trickery account does not allow us to easily accuse others of being acting manipulatively, at least not to the extent that it would if it were to define manipulative action relative to objective standards for belief, desire and emotion. The manipulator can always claim to be sincere in their beliefs, at least if they are not using inherently manipulative techniques which require a manipulative intent to be intentionally used.

This is not an unfamiliar problem. We do not have access to a person's internal mental life and as discussed agents may not even know their own intent at the time. Rather we must gather evidence about a person's intention and reason about them. Consider a racist person, who when asked denies in all cases that they are racist. They claim to love all races equally. However, they vote for candidates who explicitly state racist views, avoid communicating with those of other races, avoid where possible any interaction with those of another race, behaving dismissively towards them when forced to interact with them. In behaviour, they are indistinguishable from a person who does hold racist beliefs. What then, does their denial mean to us, when they have every reason to lie? In this current age, open racism is usually disparaged. The testimony of a potential manipulator is similarly open to doubt, as the manipulator has every reason to lie to us about the motivation behind their actions. If a person is manipulative, we can treat their actions as evidence of their intentions, as we do other actions and corresponding intentions with motivate them. The preacher has a large personal stake in what they are preaching. The more people believe there is an apocalyptic event on the horizon, the more likely she will be enriched by their purchase of her survival equipment. While we cannot say for certain they are a manipulator, we can say that they are acting how a manipulator would act, and thus the evidence lends itself to a presumption that she is one.

To conclude, we should not be daunted by the fact that we cannot know deductively that a person is manipulative in many cases. Either they are using a manipulative technique as in the cultist case, or we can examine their reasons and behaviour to make our judgment.

The Odd Manipulator Problem

The sincere cult leader counterexample has been solved by arguing that some methods of influence are intrinsically manipulative, such that they cannot be used without constituting a manipulative act. Knowingly choosing to use these techniques is knowingly manipulating the target of the technique. This introduces an element of objectivity to our application of the definition. It is not entirely subjective. Only what is considered 'ideal' is defined relative to the beliefs of the influencer. This is useful because it allows us to maintain that apparently sincere agents can act manipulatively if they employ certain techniques. Sincerity or paternalistic motivation is not an ironclad defence to the accusation that an agent is acting manipulatively.

However, we can consider another example which once again calls into question whether it is wise to adopt the relativistic approach as we do in MA3. We can now consider the odd manipulator. The odd manipulation does not share the same definitions of what is ideal that Noggle describes. Noggle presents the 'ideal settings' for rational thought, accurately named deliberative standards by Hanna[47], as governing the beliefs, desires and emotions of all agents. However, we may be able to conceive of an agent who does not share these standards. This presents a dilemma both for how the person may be a subject of manipulative influence, as well as how they may manipulatively influence others.

[47] (Hanna, 2015, p. 633)

The odd manipulator has strange or unconventional ideals by which they evaluate their beliefs, desires and emotions, or evaluate the ideal-ness of other agent's beliefs, desires or emotions. Let us use an example of an odd manipulator, Casanova.

Casanova: Casanova believes that the purpose of life is to have as many romantic relations with as many people as possible. Consequently, he thinks that a person should make themselves as attractive as possible to the current target of their romantic interest. As a consequence, he does not hold that it is ideal to have only true beliefs. Rather, he believes it is ideal to hold whatever beliefs may make a person more likely to form a romantic relationship with their current beau. To use political examples, if he meets a young socialist he finds himself attracted to, he mirrors their beliefs and is soon parroting Marxist theory. After that relationship has ended, he might meet an attractive conservative organiser, and suddenly be invested in traditional family values.

An observer to his behaviour would identify Casanova's actions as manipulative. After all, he appears to be insincere in his beliefs, merely expressing whatever his romantic target wants to hear. He is surely lying to them, deceiving them with a false picture of himself that he knows they will find attractive. Presumably, the socialist would feel she had been lied to if they saw him with the conservative months after their relationship has ended, with an entirely opposite set of political beliefs. However, Casanova believes himself sincere. He genuinely holds whatever set of beliefs he thinks his romantic partner will find attractive. It is this sincerity, or apparent sincerity, which makes him so effective at seduction. Casanova would attest that the ideal standard for his beliefs does not include any ideal for believing what is actually true, but whatever his romantic target finds attractive. Therefore, Casanova does not manipulate his partners. He does not make them less ideal in regard to their beliefs. When he influences them such that they believe he has certain political beliefs, they believe something that is true at that time.

Moreover, if he has a friend who is having romantic troubles, he will seek to convince them of whatever beliefs make them more attractive to their current romantic interest as well. If his friend wanted to start a relationship with a devoutly religious person, Casanova would influence them to believe in that religion's doctrines, regardless of if Casanova himself believes they are true. Casanova does not believe that he is making his friend less ideal, but more ideal by Casanova's own understanding of what is ideal for his friend. Thus, Casanova does not manipulate when he seeks to influence his friend to believe statements Casanova does not believe are true.

There are other examples where we can imagine agents holding their beliefs, desires, and emotions to different standards than what Noggle prescribes. Consider a party official in a totalitarian regime, who knows that for their career and survival they must believe the party ideology, regardless of contradictions or falsehoods. Whether something is actually true hardly matters to them. Of course, we could seek different explanations for the thoughts of such agents. Perhaps they wish they could believe what is true and recognise that in an ideal world they would only believe what is true. Therefore, they recognise that by becoming a party agent, they will not be ideal, but accept that they will truly internalise and believe falsehoods out of a desire for safety. Or perhaps they still believe that the ideal for their beliefs is to only hold true beliefs but pretend otherwise to be safe. These alternative explanations do not entirely rule out the existence of such agents for whom the truth is a flexible concept.

To explain the behaviour of the odd manipulator problem in accordance with our definition, we must take a stance on the objectivity of the ideals themselves. Instead of claiming there is an objective truth, or objective standard for when an emotion is appropriate etc, we must hold the standards themselves, bereft of application, are objective. One argument which lends itself to this

approach is that while Casanova and his like may appear to not operate as manipulators when applying their own ideals, they do appear to be subject to manipulation when they are the target of such influence.

Consider if Casanova was buying a car and had the bad luck to visit a dishonest car dealer. The owner lies to Casanova about the mileage the cars he is advertising have travelled, their prior ownership and their market value. As a result, Casanova buys a car for an extortionate price. Casanova appears to have been manipulated, as he believes falsehoods about his new car, which led to his decision. Certainly, the shady car dealer thinks he has successfully manipulated Casanova, although he assumed that Casanova himself held to the ideal that he should only believe what is true. I hold that even Casanova would recognise that he was lied to and manipulated. Not only should he think that the dealer acted manipulatively, but that Casanova himself was made less ideal, and thus made a bad decision. Presumably this relies on Casanova recognising that at least in some contexts, it is better for him to believe a truth rather than a falsehood. We could now hold that Casanova believes a more complicated, disjunctive ideal governs his beliefs. That would be that it is ideal to believe only truths, unless they make him less attractive to his current romantic target. If we hold this, it appears we are taking the approach of relativizing the standards for beliefs, desires and emotions, relative to the goals a person holds. Casanova is operating with ideal beliefs, if those beliefs are practical for the context of his goals. Amending the trickery definition to take into account what is in the self-interest of the influenced agent is another criticism of the trickery account which I will consider. For now, I will simply note that adopting this position introduces its own problems in regard to defining paternalistic manipulation. So, I feel comfortable dismissing this as a practical amendment to our definition.

I believe we find the solution in Jason Hanna's work. The ideals Noggle describes are objective "deliberative standards".[48] However, unlike Hanna who favours creating objective standards for what emotions are appropriate in whatever context, etc. we are only grounding the existence of standards themselves as objective. What is true or not is a matter for debate and may even be subjective or relative on some views, and agents can have reasonable disagreements on those facts. What is objective is that the agent better placed to make rational decisions, is one who believes only that which is true and attends only to those beliefs which are relevant. They do not have desires which do not correspond to their held beliefs about what they have reason to do. They feel appropriate emotions, and their emotions make salient relevant facts for the context. What exactly is appropriate can be debated, whether we should strive to feel only that which is appropriate is not. That we must hold to be objective to avoid the pitfalls of the Casanova example.

Acting irrationally may be beneficial to a person's goals, it certainly is for Casanova and the party official. However, in so far as acting and thinking rationally is a universal good, we can evaluate agents as better off in so far as they hold themselves to these standards. Casanova then, we can say is acting irrationally if he is genuinely swapping his political beliefs on a whim to make himself more attractive. Manipulative action is making a target of influence fall short of the ideals, though it is relative to the beliefs of the manipulator about the application of those ideals. Therefore, we can conclude that Casanova is still deceiving his targets, because they assume that he is acting rationally, and expressing beliefs that he holds for rational reasons. If he were to share the fact that he only holds such beliefs because he disregards truth in this context, they would question whether really holds those beliefs sincerely at all. We can then develop the account of manipulation as trickery by

[48] (Hanna, 2015, p. 635)

stating that the ideals for belief, desire and emotion are the objective standards for good reason. Those who do not hold those ideals are acting irrationally.

Manipulation, Self Interest and Paternalistic Manipulation

One great use of the trickery account is that it allows us to exclude some examples of non-rational methods of influence from being considered manipulative. This is especially useful when we want to consider influence which is clearly intended to make a person more ideal. Of course, considering the previous discussions of the sincere manipulator, we are considering influence here which is non-rational, so not persuasion, but does not make a person less ideal in one respect for the sake of making them more ideal in another. Noggle calls this "non-rational counselling".[49] One example of this is the use of anti-smoking messages and images which are printed on tobacco products by governments around the world. In the United Kingdom, tobacco products must be labeled with warnings regarding the harms of using them, including graphic images of associated health conditions. One perspective on these designs would be that they are manipulative, with the state acting to influence smokers in a non-rational way to stop smoking. However, if we view this through the lens of the account, we can argue that it does not make smokers less ideal, rather it makes them more ideal. The risks of smoking, long-term as they are, are not particularly salient to users. This is especially true for young people, who could go on for many years of smoking without facing a single consequence. The argument is that these risks *should* be in the forefront of a person's mind when they are deciding whether to not to consume tobacco products. These labels do not necessarily constitute a rational argument to those who view them, though some do. Others are simply intended to shock with graphic imagery and remind smokers of the health conditions they risk. Many smokers have the correct beliefs about the risks due to previous education. The labels make emotions about the risks more salient, and those relevant beliefs more present. We may have other criticisms to make about the images, but that they are manipulative cannot be one of them.

One could take this to imply that paternalistic manipulation is impossible, and that any case of influence by non-rational means which aims to benefit the target is a case of non-rational counselling. However, I would argue a person can be manipulated in such a way as that they benefit from the manipulation. For example, consider if I had a friend who was an enthusiastic conspiracy theorist and a heavy smoker. Worried for his health, I might influence him to quit smoking by convincing him that nicotine is a drug made by the government to dull the mind and make him more susceptible to government mind control signals. This is a lie; I am not a conspiracy theorist and do not believe nicotine is part of a government plot. Therefore, I believe I have made him less ideal, even if that led to a decision which is best for his health. Remember, by 'less ideal', we are referring to the beliefs, desires, and emotions of the target from our perspective as the influencer. We are not referring to a person having beliefs, desires or emotions which benefit them in some way.

This brings us to a criticism of the trickery account, or more specifically an argument that the trickery account should be amended to take into account a connection between manipulation and self-interest. The proposed amendment is that influence should only be counted as being manipulative if it goes against the self-interest of the influenced agent. This comes from Anne Barnhill and her paper '*What is Manipulation*'.[50] We might consider this comes from an intuition regarding practical reasoning. As stated in the discussion of the odd manipulator problem, the 'ideals' themselves are objective standards for good reasoning. However, in their application it can be beneficial to the target to have otherwise unideal beliefs, which aid in attaining the outcome the agent wants. What

[49] (Noggle, 1996, p. 49)
[50] (Barnhill, 2014, p. 52)

is unideal in general could be considered ideal for the context. In developing this point, she uses an example of a politician seeking to be re-elected:

"**Reelection II**: The President's political advisor believes that it is in President X's best interests to be reelected, but is worried that he is not sufficiently motivated. She knows that, as a graduate of Yale University, he feels an intense competitive malice toward graduates of Princeton University. In order to motivate him to do what it takes to win reelection, she reminds him that his opponent is a graduate of Princeton University."[51]

Barnhill does not view this as manipulative. We might think that feeling a childish rivalry based on university membership is not an appropriate emotion. Further we can assume the political advisor finds this childish. However, she also knows that the President's goals, to win re-election, will be advanced by him feeling this emotion. Barnhill does not think this example is one of manipulation. What motivates this intuition could be her separation of manipulation by ideal response, and manipulation by non-ideal response. Trickery which tends to promote the self-interest of the target in a given context cannot be the latter, as it actually produces an ideal response *for the target in the current context*, assuming the goal of practical reason is to promote the self-interest of the target. We would still maintain that the ideals, such as feeling only appropriate emotions, are objective standards for reasoning, but the context, especially of what is in the self-interest of the influenced person, are crucial components in changing our understanding of how they should be applied.

I do not think that we have to amend the trickery account to explain this example in a way which accords to our intuitions appropriately, nor do I accept the motivation behind it regarding practical reason. Firstly, I agree that **Reelection II** is not necessarily a case of manipulation. Although the political advisor may think that the president is being childish, she may not hold that it is inappropriate for him to experience that emotion. To manipulate, the influencer must believe they are making their target fall short of an ideal, in regard to their opinions of what would and would not be ideal for the target. This is based on their understanding of the target. While she may think that she herself would not want to participate in such childish rivalries, she may understand that the president, to her knowledge, does find it appropriate. Furthermore, she may believe that it is permissible for some people to be more childish than others. As long as the advisor believes that this is the ideal setting regarding his emotions, considering what she knows about him, then this is sufficient to say she is not acting manipulatively. There is no requirement that every ideal must be applied as if it were objective, applying to all persons equally, even if we disagree on what we regard as the objective 'correct' approach.

I am unsure if Noggle himself would agree with this tactic, but it certainly seems to be a natural conclusion to draw from the way he phrases his definitions. What indicates he may not agree is the strong desire he expresses to maintain a connection with lying. "If I try to get you to believe that *P*, when I think that not-*P*, then I lie—even if *P* turns out to be true. Similarly, I act manipulatively if I try to get you to fall short of what I think is ideal, even if I am wrong."[52] While you and I may have different opinions about the truth value of *P*, what matters if I try and convince you of a truth value opposite to what I believe is the case. However, it is not clear that the notion of appropriateness is as easily wrong or right as some facts. Opinions about truth may be something we cannot help but apply to all persons. For emotions however, it appears we can hold ourselves and others to differing standards coherently, and thus encourage other agents to feel emotions we think appropriate for them, but not for ourselves. However, there are some circumstances where we will hold ourselves

[51] (Barnhill, 2014, p. 71) my emphasis
[52] (Noggle, 1996, p. 48)

and others to the exact same standards. At the very least, there is no obstacle to understanding **Re-election II** through this lens.

Secondly, Barnhill wants to connect manipulation and self-interest more closely, motivated I think on the understanding of the purpose of practical reason. I believe we are better served looking at practical reason not as a way to promote self-interest, but make rational decisions based on whatever hypothetical objective an agent has. Consider this modified version of **Re-election II**.

Reelection II(M): The President's political advisor believes that it is in President X's best interests to be re-elected but is worried that he is not sufficiently motivated. She knows that, as a graduate of Yale University, he feels an intense competitive malice toward graduates of Princeton University. In order to motivate him to do what it takes to win re-election; she lies to him by telling him that his opponent is a graduate of Princeton University. He actually graduated from another university."[53]

In this case, it is still in President X's self-interest to believe that his opponent graduated from Princeton University. However, it is still clear that the advisor lies to him. This might be a justified lie on balance, or a lie he would thank her for later. However, it remains a lie. The goal of practical reason is to make a rational decision, and to do that most effectively we aspire to have ideal sets of beliefs, desires, and emotions in the way the trickery account proposes. It can be in a person's interest to be manipulated, and still somehow harmful to them if you believe that manipulation is always morally dubious. Nevertheless, it remains manipulative. For these reasons I reject the idea that the theory can be improved by amending it to explicitly reference self-interest. Consequently, it stands that paternalistic manipulation can exist.

In summary, the trickery account does not need to be amended to define manipulation as only occurring when being made less ideal harms the target's ability to pursue their self-interested goals. Manipulation can be carried out for paternalistic goals. It is an open question whether or not such manipulation is morally problematic or whether it would hypothetically be consented to by the target. What matters at the moment is that MA3 does not need to be altered further and that as a consequence of rejecting this amendment, we are committed to the idea that paternalistic manipulation can occur.

Covert Manipulation, Agent Authorship and Temptation

The final criticism of the trickery account I want to consider comes from the paper *'Online Manipulation: Hidden Influences in a Digital World'* written by Daniel Susser, Beate, Roessler and Helen Nissenbaum. Hereafter, I will refer to them as Susser et al. Their paper develops a different definition of manipulation which they proceed to apply to digital environments. In the process, they present two criticisms which are applicable to the trickery account. The first is that it allows for non-covert manipulation, which they argue is impossible. Second, they argue that it "fails to capture what is distinctive about manipulation – that it undermines our sense of authorship over our decisions."[54] Of course, some of their arguments are based on contrasting the trickery definition with their own definition.

Essentially, they argue that if an instance of influence is persuasive, then the agent being influenced is faced with argument or other persuasive techniques they choose to accept or reject. If they are coerced, then they still face a choice. If faced with an irresistible threat, such as a gun to the head, then they still have the choice, albeit a heroic one, to resist and choose to be shot rather than give in

[53] I have named it Reelection II(M) as Barnhill names another example as 'Reelection III' in her own paper. I do not wish to cause confusion in what example I am referring to.
[54] (Susser, Roessler, & Nissenbaum, 2019, p. 18)

to the influence. Susser et all argue that manipulation is distinctive because it robs the agent of such a choice. This is the approach of manipulation as bypassing reason, expanded to bypass proper choice making by the agent. For them, manipulation occurs when an agent is "infected by external machinations".[55] By necessity, manipulation must be covert. If you become aware that somebody is trying to influence you, then you regain your ability to accept or reject their influence. You are essentially immune to manipulation once you become aware somebody is trying to perform it.

Susser et al use an example of a group of friends to demonstrate an example of what they think is manipulative action which the trickery account does not include. Moreover, it would be included under their definition.

Binge Drinker: Teddy is prone to drinking heavily in small bursts. Teddy plans to spend the night browsing the internet with a bottle of whisky and tells his friends about this. Understandably, Teddy's friends are concerned about his habit and conspire to share articles on their social media pages about the dangers of drinking. They are aware that as a voracious social media consumer, Teddy will see the articles, glass of whisky in hand. When Teddy reads the warnings about the dangers of binge drinking, its effect on his health and its high calorie content etc, Teddy decides to stop drinking for the night. [56]

Susser et al claim Teddy's friends have manipulated him, as "your friends have insinuated themselves covertly into your decision-making process and redirected it to their own ends".[57] They argue Teddy has made a more ideal decision by means of the influence, deciding not to drink heavily. We have already explored amending the trickery account to consider self-interest and rejected it. Rather, the trickery account identifies the cause of manipulation as making an agent fall short of an ideal governing their reasoning, rather than causing an agent to make decision which is or is not in their interest.

It is true then that the example would not be manipulative under the trickery definition. The friends share articles which we can assume are sincere, and assume they are beliefs the friends themselves share, considering they are trying to convince Teddy to stop binge drinking for the sake of his health. An agent who reads the articles is influenced in a way which makes them more ideal from most people's perspectives. Neither the friends nor the authors of the articles presumably have manipulative intent. In fact, we could call this covert persuasion on the part of the friends, overt in the case of the article writers.

It's also true to say that Teddy's friends have covertly interfered with Teddy's decision-making process. Why is this notion of covert interference essential to manipulation in the view of Susser et al? It is because they assert manipulation is unique as a method of influence because it robs an agent of authorship of their own actions. What they mean by this is "to displace them as the decider"[58], to bypass their conscious decision-making processes and reduce the control that they had on their own actions.[59] If you are an ideal self-governing agent, decisions are made using your thought process, your desires, and your ends. In the example of the binge drinker, only the last is potentially suspect. Teddy saw articles, thought about the arguments they presented, considered his desire to drink, and decided his competing desire for a healthy life took priority over his desire to get drunk. The assertion would be that Teddy lacks ownership because in making this decision, he

[55] (Susser, Roessler, & Nissenbaum, 2019, p. 19)
[56] (Susser, Roessler, & Nissenbaum, 2019, p. 19) paraphrased heavily.
[57] (Susser, Roessler, & Nissenbaum, 2019, p. 19)
[58] (Susser, Roessler, & Nissenbaum, 2019, p. 16)
[59] (Susser, Roessler, & Nissenbaum, 2019, p. 16) footnote 61

served the ends of those who first presented him with the influence. They affected him without his knowing in pursuit of their own ends. Therefore, while Teddy retains a large amount of his ownership of the decision, it is less so because he was made subject to the influence. His self-governance has been partially tarnished.

I dislike this argument. All influence is interference, all intentional influence is intentional interference. Moreover, I am not convinced that whether influence is covert is sufficient for it to rid the target of the influence of all ownership of their decision, nor transform otherwise persuasive influence into manipulative influence.

Consider a simple piece of influence, a sign in your workplace's restroom reminding employees to wash their hands. Imagine you see this sign and wash your hands whereas otherwise you would have absentmindedly left without doing so. It has influenced you and done so quite transparently. The sign itself was not covert influence. You ponder that you hadn't seen the sign before. Perhaps it was put up recently by your employer as part of a hygiene campaign or to comply with some new law or regulation. However, unbeknownst to you, it was actually placed by a hygiene obsessed colleague or yours. Moreover, it was targeted at you in particular, as they witnessed your dirty habit of not watching your hands and felt disgusted. If the sign was generic, it would not be manipulative. It does not make your decision-making process which led to your actions non-ideal in any way I can find. I also do not think it is manipulative if your colleague was to blame. All that changes is the motive behind the influence, and the opaque nature of such motives. This does not change how the influence has operated upon you. I do not think I would feel manipulated if I knew their motives either. Unfairly targeted, perhaps. Maybe I would feel annoyed that my colleague went out of their way to interfere in my affairs. However, I would feel similarly aggravated if a colleague disagreed with my political beliefs, and constantly tried to change my mind through persuasion. All sorts of influence can be unwanted, whether the influence be persuasive, coercive or manipulative. Therefore, just the fact that influence is covert, does not appear to change the way the influence operates enough to make this the sole feature which makes an act manipulative. Neither is it a unique characteristic persuasion and coercion do not share.[60]

Return now to the example of Teddy the binge drinker. Why did Teddy decide to stop binge drinking in their example? Presumably for the end of his health, convinced as he was by the articles. If his friends share this end, what of it? Presume they had a different aim. Perhaps Teddy is a horrible drunk and they don't want him bothering them when he drinks alone, causing mischief on social media. I would assert that Teddy has been used. He served their goals as well as his own. This is something Teddy may be entitled to object to, but once again being used is present in cases of persuasion. Susser et al equate being used with being manipulated. I think they are distinct, but overlapping concepts. If I persuade my children to go to bed by reminding them that they have to get up early tomorrow, I do not manipulate them merely because I also want to watch my television show without their presence. Yet to manipulate somebody to serve your ends is to use them as well.

I have argued that **Binge Drinking** is not a good example of manipulation. I have also used it to argue that simply because influence is covert, this is not sufficient to also claim that an agent's authorship

[60] Of course, there is a difference between communication—and therefore influence—which intrinsically carries the assertion of a particular sort of intent, and only not mentioning intent as part of the communication. We can cast our minds back to the waiter in Vegetable Soup, who I asserted, by stating the soup was suitable for those with a vegan diet when he was unsure, communicated that he was sure, and thus deceived the customer. Not all influence also communicates associated, implicit assertions. Exact intentions behind a communication are not necessarily always asserted as part of a communication.

of an action is in doubt. However, I have yet to show why I think that it is not essential amend MA3 to consider the notion of agent authorship.

Temptation is a phenomena Susser et al do not think should be considered manipulative due to their definition.[61] The trickery account definitely includes temptation. They use an example similar to **Peer Pressure**, so for the sake of clarity let us use that one in place of Susser et al's, rather than introducing another example which is nearly identical.

To restate the example, Sarah is trying to cut down on her drinking while her friend Percy thinks she is being a stick in the mud and should continue her drinking habits. Therefore, he buys Sarah drinks to tempt her to break her word and return to her bad habits. On the trickery account this is a classic case of influence intended to make a person less ideal. Sarah desires to stop drinking heavily. Percy enflames the competing desire to drink which if successful, would cause her to be in a state of motivational akrasia, with no motivation to fulfil her desire to stop drinking, which she has reason to do.

Susser et al argue that this cannot be manipulative, as Sarah retains authorship of their own actions, and the influence is overt. If Sarah is not a hopeless alcoholic, then surely she can resist the urge to drink. If she cannot, this would be an irresistible compulsion, she would have no choice, so it would be coercive for Percy to place a drink in front of her. If she can resist it, then it was Sarah who made the decision to drink in "all [her] complexity"[62]. They argue she was persuaded with bad reasons.

I have two issues with this criticism. Firstly, consider if Percy was covert. Consider if he asked the bartender to announce that Sarah were the winner of a lottery of regulars and thus were entitled to some free drinks, which were brought to the table. The situation is otherwise the same, Sarah is tempted by the drinks. By their own metric in the binge drinking example, it is now a manipulative situation, as the influence is covert. However, I do not see how this would affect Sarah's ownership over her decision any differently. She still makes the decision, in all her complexity, even if she was unaware that this was a plan of Percy's. Again, covertness does not affect self-governance in the way Susser et al want it to.

Secondly, temptation is more of an interference that Susser et all want to recognise. What is important for a decision being under the ownership of an agent, is for them to have the opportunity to evaluate the choice. Susser et al are arguing that if influence is overt, then an agent can absorb that influence into their existing process. However, the point of manipulative influence is that it is this decision-making process which it aims to disrupt. Consider gaslighting, the practise of causing somebody to doubt their beliefs unjustifiably. An abusive partner may insist that you were flirting with a person at a party, despite the fact you believe you did not. They will repeat it and repeat it, causing you to doubt even beliefs you may have been sure of. These techniques are effective, regardless of your awareness of them. I can know somebody is attempting to gaslight me but may not be able to help the doubt that creeps into my mind, putting me in a non-ideal state of belief which can affect my decision making. Temptation functions in the same manner.

It is worth noting that Susser et al's conclusion, if it were true, would lead to what I would say is an underinclusion of relevant phenomena. Specifically, those used in abusive relationships. To repeat, Susser et al conclude overt manipulation is impossible, as "either you can resist the influence and

[61] (Susser, Roessler, & Nissenbaum, 2019, p. 21)
[62] (Susser, Roessler, & Nissenbaum, 2019, p. 19)

have therefore been coerced, or you can resist it and do not, in which case you have simply been moved by bad reasons."[63]

For those in abusive relationships, they are either powerless victims or weak people, who are either cowed entirely to the point where they have no choice otherwise, or moved to action by the bad reasons presented by the abused. I do not think this gives sufficient merit to the complexity of the situation individuals find themselves in. For example, to be made to feel guilty for something you have no reason to feel guilty for is not a choice a person makes. One does not make a decision to feel guilt. Yet that guilt may lead an individual to explain away violence or abusive behaviour. After all, you did something wrong, so you deserve it. The rational decision-making process is purposefully interrupted by manipulator or abusers. When free from the situation and with a clear head, survivors of domestic abuse may indeed look back in horror at the decisions they made, but they were not fully authors of those decisions. They had been manipulated.

To summarise, Susser et al propose that manipulation is covert influence which consequently robs targets of the ability to transparently evaluate the influence using their own rational capacities. Temptation is usually not covert, so Susser et al argue that what is categorized by the trickery account as manipulative action is instead persuasion using bad reasons. I have argued this is not the case. I've argued that covertness is not necessary for an act to be intuitively manipulative, and that interfering in another agent's rational process does not necessarily rob them of their ownership of an action. Moreover, I argue that manipulation such as temptation involves not just bypassing the ability of targets to transparently evaluate overt influence, which is used against them, but subverting it by causing mistakes in this process. A lack of ownership over our actions may indeed be extremely relevant to the ethics of manipulation, but I do not believe the fact that MA3 lacks a focus on it causes any over or underinclusion of relevant examples. Therefore, in my view it is not necessary to alter the definition. It is sufficient here to conclude that manipulation as trickery does capture this concept, as manipulative influence interrupts the decision-making process of an agent. This need not be covert. Manipulation such as temptation, can be blatant, yet still effective.

Summary

In this chapter I have explained and developed Noggle's account of manipulation as trickery. From the starting point of defining manipulation as influence which is intended to bypass or subvert our rational capacities, the trickery account gives us a structure by which to understand the variety of ways manipulation functions. Manipulative actions are those which are intended to make us fall short of one or more of the ideals by which we govern our beliefs, desires and emotions. I depart from Noggle's trickery account by understand manipulative intent as a sort of intent which can be mostly obscured at the time to its owner, yet still the motivating force behind manipulative actions. Therefore, I have argued that people can manipulate recklessly, with a reckless intent as well as a purposeful, planned intent. The definition of manipulation as trickery I propose is:

*MA3: An act of intentional influence directed at an agent, where the actor acts **without** the intention that the agent **does not** fall short of one or more ideals which govern their beliefs, desires or emotions, relative to what the manipulator believes are ideal for the target agent.*

This trickery account was still subject to much of the same criticisms which have been levelled against Noggle's original account. I have considered what I believe to be the strongest objections to the trickery account as definition, and I believe I have adequately dealt with them all. Now, it is time

[63] (Susser, Roessler, & Nissenbaum, 2019, p. 21)

to turn to an ethical analysis of this account. As we now have a definition of the feature of manipulative actions, we can consider which of those features are ethically important.

Chapter Three: The Ethics of Manipulation

Manipulation is wrong. How could anybody think otherwise? To be labelled a manipulator is to be called a bad person, somebody who should not be trusted. It would be a strange sort of person who regarded engaging in manipulative behaviour as a positive exercise. Even if a person did embrace the label, we might think they are doing so ironically, or otherwise in recognition that it is wrong. Perhaps they have embraced the role of the villain and take pride in being "bad". If not that, they may recognise that manipulating is taken to be intuitively wrong but repudiate the conclusions of common-sense intuitions as unsophisticated or believe themselves above such notions.

Perhaps our villain is right to dismiss the conclusion that manipulation is wrong. Just because a concept has negative connotations, it does not mean it is wrong. To call somebody a killer is also to label them with a pejorative term, but police officers and soldiers can be killers too. Not all killers are murderers. A useful example is one of hostage negotiation. Imagine a terrorist has taken an innocent person hostage. They threaten to kill the hostage unless their political demands are met, in this case releasing their comrades in arms from government custody. Barring exceptional ethical considerations[64], we might assume any means of influence is permissible if they result in the terrorist letting the hostage go free. Presumably, if it is permissible to send in a crack team of law enforcement who will shoot the terrorist dead to prevent harm to the hostage, it is also permissible to manipulate them. Furthermore, the skills of manipulation, in this case possessed by the hostage negotiator, are useful skills for them to have. We could view a hostage negotiator without the ability and the will to skilfully manipulate those they speak with a poor hostage negotiator, unfit for the position.

We might further this point with less exceptional cases. Our examinations of manipulation in the previous chapter led us to conclude that deception, in all its forms, is inevitably a form of manipulative action. To be a liar or a deceiver is hardly an ameliorative term either. Yet I assume comparatively few would suggest that lying is always impermissible than would accept that manipulation is always impermissible, though some philosophers obviously do take such a position against deception and have expansive and sophisticated reasons for doing so.

Regardless, this chapter will move beyond the simplistic intuition that manipulation is simply wrong. As we have explored in the first chapter, we distinguish between different methods of influence because they are interesting. The moral status of manipulation as a method of influence is easily its most important feature and I would suggest the primary reason the topic draws our interest. In this chapter I will examine the trickery account to see what, if any wrong-making features are present in manipulation as I have defined it. We can consider if there are any instances of manipulation which lack these features, and thus whether or not cases of manipulation can be innocuous.

[64] By "exceptional ethical considerations", I am referring to such circumstances as the hostage negotiator being a family member of the terrorist, so perhaps on some views having exceptional moral duties towards them. For the purposes of illustrating my point with this example, I assume that if a great harm or wrong can be justified, any lesser harm can presumably be justified too. I'm sure there are cases where this is overly simplistic, but I do not want to consider them here. I also assume here the terrorist isn't fighting for a just cause (e.g., the end of a tyrannical and terrible government which is committing a genocide), which might make their actions permissible and attempts to stop them not so.

The Wrong-Making Features

A wrong-making feature of an act is the feature which lends negative moral weight to that action in a given context. In other words, the presence of wrong-making features are what make an act wrong. Some acts have multiple wrong-making features, some of which are part of the act itself, others which are part of the context the act takes place within.

It is the former type of wrong-making feature we're interested in. Manipulation can be wrong, like many if not all acts, because of the context it takes place in and the consequences the act has. Let's take **Ticket Deception** as an example. I'll restate it here for convenience of reference:

Ticket Deception: Mr Fisk is an important businessman. He abuses the ticket salesperson at the train station when the machine which prints the tickets malfunctions. The salesperson purposefully tells Mr Fisk that his train is leaving from a different platform that it actually is. The salesperson gets on the wrong train and ends up late for an important business deal.

The fact that Mr Fisk taking the wrong train means that he was late for an important business deal is a feature of the ticket salesperson's manipulative act which is wrong-making, but only in the context of this particular act, not manipulation in general. The lie might have quite catastrophic consequences. For example, imagine if Mr Fisk was negotiating a rescue deal for his failing business. Being late might sour negotiations to the extent that all of Mr Fisk's employees lose their jobs, causing a cascade of misery, all because of one lie. The point is that often when we think of examples of manipulation, we may be tempted to point to the negative consequences of a particular act as the wrong-making feature of it. They are certainly part of what makes an individual act wrong, but they are not wrong-making feature of manipulation as a category of influence. However, I am not taking an anti-consequentialist approach either. It may well be that the wrong of manipulation is in a consequence which, because of how manipulation is defined, always follows necessarily whenever a manipulative act takes place. I only mean to highlight here that we aren't interested, primarily, in the consequences, even catastrophic ones, that acts of manipulation can cause due to the circumstances they take place in. Rather, in order to evaluate whether manipulation, as a type of influence, is wrong, we must look to wrong-making features which are present *because* the act is manipulation, rather than another sort of influence.

So, to act manipulatively is to intentionally try to influence another agent where the influencer acts without the intention that their influence does not cause the target to fall short of one or more of the ideals which govern the agent's beliefs, desires or emotions, relative to what the influencer believes is ideal for the target. What wrong making features are present in all acts which meet this definition, if any?

Noggle gives the following simple feature for his trickery account. It would also apply to our amended trickery account, with only slight additions. Rational agents have a clear interest in attaining the standards which govern their beliefs, desires, and emotions. These are integral to how a person operates as a moral and rational agent. Therefore:

"Acting manipulatively, toward someone, then, is an affront to her as a rational and moral being; for it is an attempt to thwart her moral and rational agency, which has as its goal the correct adjustment of her psychological levers. To attempt to thwart the goals someone has qua rational moral agent is to fail to respect her rational moral agency. And since a person's rational moral agency is crucial to

her personhood, to fail to respect it is to degrade her; it is to treat her as less than a person. And for that reason, it is wrong."[65]

Let's consider this step by step. An agent is someone who can form and pursue goals. Part of forming and pursuing goals effectively is to try and fulfil the ideals for belief, desire, and emotion we have described. Noggle calls them "goals someone has qua rational moral agent". What he means is that these are 'core' goals, which one has in the capacity of being a rational moral agent. Agents are beings which form and pursue goals, and therefore, seek ideal psychological traits as part of forming and pursuing goals effectively.

Moral agency is integral for an entity to be a person, and thus qualify for a certain respect from others. To attempt to thwart this process then, by influencing them to fall short of one or more of the ideals, is to fail to treat a person as a worthy agent. This is degrading the agent. This is what Noggle identifies as the intrinsic wrong-making feature of manipulation.

There is a logical step missing, we might think, from depriving an agent of the full capacity to act as an agent, and this being necessarily wrong. Perhaps this presumes a right to the full pursuit of agency. If we make this assumption, the wrong of manipulation is easily explained. As agents, we have a right to the full pursuit of agency, and others are obligated not to impede us in this, all things considered. However, Noggle instead focusses on the 'disrespect' shown by manipulating a person. If we do not want to make an assumption about the existence of a such a right, disrespect might also give us a wrong-making feature.

Disrespect is not always associated with wrongdoing. Rather, it is often associated with social convention. For example, one society may hold that in order to respect the deceased, offerings should be made to the deceased on specific holidays. Others believe that to respect the dead is to exhume them and involve the corpses in the festivals on those holidays.[66] This latter ritual may leave those of other cultures aghast at what they perceive as disrespectful desecration. Of course, these things are culturally relative. When we talk of respect then, we are really stating that something is owed to a person. When we say a person 'deserves respect' they are owed a type of behaviour towards them, an ascription of value which rationally entails being treated as something with that value, whatever that means in context. When a person has been 'disrespected', they are not being given the behaviour they are owed, or being ascribed the value they deserve and hence treated as something with that value. For Noggle then rational agents have a value which demands we behave in a particular way towards them, and because they are rational agents, that owed behaviour is not intentionally thwarting the processes which allow them to act as a rational agent should. For us, we might amend this slightly to say that the owed behaviour is the intentional *avoidance* of thwarting their goals qua rational agent.

When considering the question of rights and disrespect, I was drawn to consider the practise of lobotomy. A lobotomy is a medical procedure where parts of a person's brain are altered or removed, with the goal of curing mental illness. Lobotomies are now a mostly shameful part of the history of mental healthcare. They were often performed without consent and on vulnerable people such as women and ethnic minorities. A woman who was depressed or anxious could be given a lobotomy to make her more able to fit her societal role, rather than as a real solution to the problems she faced. A lobotomy permanently removes the capacity for some mental activity. We

[65] (Noggle, 1996, p. 52)
[66] (British Broadcasting Corporation, 2008)

might imagine that these mental activities are required to meet the ideals. One daughter remarked of her mother, who had a lobotomy, that she "developed into a pretty reliable dishwasher."[67]

The lobotomy is horrifying. It has taken centre stage in the media in the horror genre such as '*One Flew Over the Cuckoo's Nest*'. Manipulation, like a lobotomy, practically effects an agent's capacity to act as an agent does. It goes without saying that it is less significant and usually temporary. These practical effects are significant, but the attitudes of those who inflict them are more important. As Noggle argues, the manipulator, like the lobotomiser, robs a person of a fundamental human dignity. To lobotomise a person is to deny that a person's projects, perspectives, goals, and capabilities as an agent were worth preserving. They were not worth the trouble it would take to fully engage with them. To do this is to deny that they are like you, a human agent deserving of acting as an agent. Rather, they must be something lesser. The manipulator is expressing the same arrogance and distain. The victims of a manipulator are lesser in his eyes, a necessary sacrifice in respect of the greater good. It does not matter if the manipulated believe what is false as long as they act how the manipulator desires as a result. The goals and projects of the victims are made subservient to those of the manipulator. This is to treat the victim disrespectfully, in regard to their status as a rational agent in their own right.

I believe this is the wrong-making feature of manipulation as trickery. Manipulative acts, defined as we have defined them, can only be directed towards fellow agents. By virtue of being rational agents, they have a value which others are obligated to recognise, and therefore treat them accordingly. To disrespect moral agents is to treat them in ways which do not recognise this value, instead treating them as a mere means to an end, a diminished person.

We should distinguish between a disrespect of a person's agency, and disrespect directed towards what they are pursuing with that agency. I do not mean to say that others' beliefs, as an example, must be accepted unquestioningly lest we treat them as an agent ought not to be treated. Instead, it is that capacity to hold and come to beliefs, and seek to hold only true and relevant beliefs, which it is a wrong-making feature of an act that it prevents. Of course, this wrong-making feature is present in all manipulative acts according to our definition of manipulation.

I should note here that there is a slight nuance between Noggle's account of the wrong making features of manipulation, which follows solely from his definition, and my own. In the previous chapter I argued that a definition of manipulation as trickery needed to take into account reckless manipulation, where the manipulative agent does not have an explicit intent to manipulate, but rather lack's the intent not to. Some agents do not care if they make the target of their influence more or less ideal, only that they get what they want. This best captures the behaviours of some of what we might consider the worst manipulators, such as agents who are complete narcissists, and do not consider other people as proper agents at all.

For Noggle's definition, we owe people a negative obligation. That is, they have a right which compels us to not act in a particular way towards them. They have a right to not be manipulated. However, on my definition, this is probably better formulated as a positive obligation. Agents have a right to be treated respectfully, and this means we are obligated to actively avoid manipulating them. For Noggle, it is enough to not have the intent to manipulate. For me, I have defined the intent such that a person must have an active intent not to act manipulatively when they influence somebody. This difference is a subtle one, but one worth noting.

[67] (Raz, 2014)

The Moral Status of Manipulation

I argue that the wrong-making feature of manipulation is that it results in the targeted agent being treated disrespectfully. This wrongs them, as it is treating them in a manner which does not recognise the value they have, and in fact treats them as if they have a lower value, thus degrading them. Of course, an individual manipulative act can be impermissible because of consequences unique to that particular act, but those wrong-making features may be unique to that act alone, not manipulation as a category.

Having identified this as the wrong-making feature of manipulation, we can take three approaches towards the moral status of manipulative actions. We could take an absolute position on its moral status, asserting that manipulative actions are always wrong. Consequently, we may also take the position that it is always impermissible to engage in manipulative acts. I believe this is the wrong position to adopt. Clearly, from examples such as the hostage negotiation above, it can be permissible to engage in manipulation. Moreover, we may even think that in some circumstances a person may be obligated to engage in manipulative action, should the stakes of not doing so be high enough. This opposing position would be to say that manipulation is always a *pro-tanto* wrong. What I mean by this is that the fact that an act is a manipulative one, and consequently the act has manipulation's wrong-making feature, provides meaningful moral reasons against engaging in the act. This means that manipulative acts are always wrong, but the pro-tanto reason can be outweighed by other considerations, meaning manipulative acts are permissible. This is the position that I will argue for in regard to manipulation. I argue manipulative acts are always wrong, in that they necessarily mean wronging another person. However, sometimes wronging another person is necessary or morally obligatory. Ergo, while the moral status of manipulative actions is that they are wrong, they can still be permissible.

The final approach we can take is that manipulation is *prima facie* wrong. By this I mean that while the fact that an act is manipulative is still a pro-tanto reason against the act, it is not always so. Once resolved, it does not just mean that the act is permissible, albeit still wronging a person, but that it does not wrong a person in that circumstance. This would be to say that in some circumstances, manipulation can be *innocuous*, that is having no moral consequence. In those circumstances, the fact that an act of influence is manipulative would not be a pro-tanto reason not to carry the act out.

I believe that I have already presented sufficient argument in the previous chapters as to why I do not agree with the absolute approach. Therefore, I will not dedicate further space to it. Rather, I will consider in more detail arguments that others have made which assert that manipulation can be innocuous. The difference is a subtle one, but I believe it is crucial to consider whether we can act manipulatively and wrong nobody, or whether somebody is always wrongs, regardless of whether on the greater perspective the act was still justified. Hence, let's consider some of these perspectives.

Baron and Strawson

Marcia Baron, in her paper and lecture '*Manipulativeness'*, considers manipulation as an Aristotelian vice. While she is not examining manipulation using the definition of manipulation as trickery, she examines the moral status of manipulative actions in detail. She characterises a manipulative person as "too ready to think it appropriate … to orchestrate things so as to lead others to act as he wants them to; and in those instances where it is not inappropriate … the manipulative person is too ready to employ means that should not be employed".[68] Of course, this does not rule out that it can be

[68] (Baron, 2003, p. 48)

appropriate in some contexts, only that the manipulative person is too ready to assume this. While obviously this is not our trickery account, it is not an incompatible vice. In discussing what circumstances it may be appropriate, and what means are appropriate in those contexts, Baron provides examples of what she deems innocuous manipulation, where it is appropriate to manipulate a person and uses no techniques which are problematic in the context of the example. These examples would also serve well if we examined them using the trickery account.

Baron, in a separate paper '*The Mens Rea and Moral Status of Manipulation',* gives us two reasons to think that some manipulative acts must be innocuous. The first is a list of intuitively innocuous instances of manipulative behaviour. One example would be a real estate agent scenting a house for sale with vanilla before opening the house to the public for viewing by potential buyers.[69] For Baron, the scent is a sort of non-rational influence which could be viewed as manipulative. The scent makes the buyers feel welcome, relaxed, and more predisposed to view the property in a favourable manner. Baron argues that this is "not morally objectionable at all"[70], and groups this with other "sales techniques" or "mood enhancers" such as playing upbeat music in a shop to encourage spending, or placing particular products in a shop where they are more likely to be sold.[71] Whether the latter two of these techniques should be considered manipulative under the trickery account will be considered in the final chapter, where I will consider whether nudges count as manipulative under the trickery account. However, in regard to the first example of a pleasant scent, I feel comfortable arguing that under the trickery account this is not a case of manipulative action, presuming the agent's intent is sincere. The pleasant scent will aid viewers in imagining the house as a prospective home. Arguably then, it could make a person more aligned with the ideals for belief, desire and emotion, and so not be an example of manipulative action. It would be as much a manipulative act as tidying the house before presenting it to viewers or furnishing it beforehand so as to demonstrate what rooms may look like in use. None of these are particularly intuitively manipulative. Therefore, I think this example doesn't give us good reason to believe manipulation can be innocuous.

The second argument is much more interesting and important. Baron draws from P.F Strawson's essay '*Freedom and Resentment'*. Strawson is concerned with the issue of free-will and determinism and how we may treat people if we subscribe to the latter viewpoint. He contrasts two different attitudes we can take towards people. The first is the range of involved attitudes, which we take towards those we interact with in the context of a social, interpersonal relationship. The second is the 'objective attitude'[72] To take the objective attitude towards a person is to see them as "an object of social policy; as a subject for what, in a wide range of sense, might be called treatment as something certainly to be taken account, of; to be managed or handled or cured or trained; perhaps simply to be avoided".[73] We take this attitude towards those we see as having diminished agency. Strawson states "the agent was himself; but he is warped or deranged, neurotic or just a child".[74]

Baron gives an example of a supervisor who is so arrogant as to dismiss any idea his subordinates suggest unless he believes that he himself thought of the idea. Baron argues this person is so unreasonable that given we cannot 'avoid', 'cure' or 'train' him, a person is justified in 'managing'

[69] (Baron, The Mens Rea and Moral Status of Manipulation, 2014, p. 114)
[70] (Baron, The Mens Rea and Moral Status of Manipulation, 2014, p. 114)
[71] (Baron, The Mens Rea and Moral Status of Manipulation, 2014, p. 114)
[72] (Strawson, 2008, p. 9)
[73] (Strawson, 2008, p. 9)
[74] (Strawson, 2008, p. 9)

them.[75] A person with such a boss may trick them into 'coming up with' an idea that they have been suggesting subtly to the boss for a while. The boss is deceived into thinking this was his or her idea all along. This could be interpreted either as a special feature of the situation which justifies us degrading the manager, or alternatively as a reason that we do not degrade them at all. Baron appears to argue for the latter as through their actions, they are a person to be 'managed' and deserve to be treated as such. Therefore, we do not degrade them by doing so. As Baron puts it:

"One reason, then, why manipulation is in some instances not wrong is that the person with whom one is interacting is, in those circumstances, deaf to reason. Even so, manipulating the person generally leaves a moral residue. It is something to be avoided. "[76]

By 'moral residue' I she is referring to the fact that the situation is still regrettable. We wish the person was not so unreasonable. We wish they were different, such that we did not have to resort to manipulative acts to interact with them productively. Nevertheless, it is right we do so. It is dirty work, but not impermissible. Compare it to deciding how to dispose of a broken appliance or worn-out clothing. We may find it preferable to recycle it, but if there is no other option, we commit no wrong by simply throwing it away.

The argument is nuanced. A defender of my point of view, that manipulation can, and often is, made permissible by the context of the act, does not need to worry about Baron's argument here if manipulation remains regrettable, but is justified in the circumstances. If a person is being entirely unreasonable, and as a consequence of their unreasonableness, there will be dire consequences, then a person is justified in 'managing' them to whatever extent is necessary to avoid those dire consequences. I might have moral obligations not to manipulate the unreasonable person, but I have other obligations to other people as well. This coheres with the view that manipulation is at all times morally wrong, but can be justified in contexts where it is the lesser evil. I wrong the unreasonable person, but all things considered, it was not impermissible for me to do so.

Baron is arguing there are cases where people have "done something that renders him or her a suitable subject for a dose – though only a small dose, barring the addition of further details – of what Strawson calls the "objective attitude""[77] and this is innocuous. Baron does not coach her discussion in that of rights and obligations, but if one considers agents to have a right not to be manipulated, it is as if by their conduct, they have forfeited that right. It is these cases which form a challenge to my own view. In these cases, we don't disrespect the individual, in fact, we're treating them how they deserve to be treated in the circumstances.

What immediately comes to mind when considering this point of view is that we do not usually treat persons who are not susceptible to our persuasive charms as having something wrong with them. Nor do we treat them as if they have committed some offense against us which means they have forfeited a presumed right to certain treatment. This is especially true when something as nebulous as "being reasonable" is our measure for deciding whether we are justified in manipulating an agent. Just because I believe I've made a wholly convincing argument does not mean other agents are bound to share my position. While certainly it's possible that agents can be entirely deaf to reason, we are not usually in the privileged position to declare those who oppose us as "unreasonable" and therefore claim they deserve to be treated as less than sufficiently competent agents. Baron emphasises this herself, stating that pragmatically it may be best to avoid informing people that

[75] (Baron, The Mens Rea and Moral Status of Manipulation, 2014, p. 117)
[76] (Baron, The Mens Rea and Moral Status of Manipulation, 2014, p. 116)
[77] (Baron, The Mens Rea and Moral Status of Manipulation, 2014, p. 117)

manipulation can be innocuous in these circumstances. People may seize upon this as justification for manipulating agents who merely disagree with them, rather than being deaf to reason. However, while we can take a position as to the practicality of identifying just when agents are deaf to reason, or whether we should even promote engaging in such identification, this does not affect the theoretical point being made. That it can be innocuous to act manipulatively.

I believe we can consider two cases. The first is Baron's arrogant supervisor. The supervisor is 'deaf to reason' only in so far as they refuse to consider opinions other than their own. In such cases, Baron surmises what we need to consider when deciding whether to be manipulative.

"a significant factor separating morally permissible from morally impermissible is whether our treatment of the person as someone to-be-managed is in fact called for. Is the goal so important and is the likelihood that the person can be convinced so slight as to warrant manipulating the person?"[78]

There is a tension between two ideas here. The first is that an agent being unable to be effected by reason justifies, and in fact makes morally innocuous, types of influence against them which otherwise would be prohibited. The second is that the goal of our influence must be important enough to justify the conduct.

At least for the trickery account, if not for Baron's own understanding of manipulation as an Aristotelian vice, I do not believe we need both. Either manipulation is justified based on the relative importance or value of the results of the influence, which I would say is perfectly compatible with the idea that manipulation is always involves wronging another agent, or it is justified by that agent's conduct or otherwise that agent's characteristics and is not a wrong at all in that context. We do not need both.

Ultimately, in regard to the arrogant supervisor then, the only relevant question which remains is whether the sort of traits ascribed to the supervisor really mean that we can take the objective attitude towards them, and without moral consequence wilfully ignore their agency. I do not think this is the case.

I would argue that even if we presume the supervisor couldn't be presented with any reasons by his co-workers which would convince him to change his view, this does not mean he has lost his capacity to reason in a relevant sense. Rather, he either has a different position, conscious or not, on how he holds his beliefs; alternatively he could simply not be applying the standards of reason he holds. Whichever is the case, he is acting irrationally. He is not changing his beliefs so that he only believes what is true and relevant. Rather he is seeking to believe whatever satisfies the demands of his ego. This fact in of itself does not justify lying to him, or diverting his attention to irrelevant beliefs, etc.

Consider the following example:

Superstitious Habits: Mary holds some beliefs in superstitions. One is that you should never cross paths with a person on a set of narrow stairs, as it brings bad luck. John believes Mary's beliefs are irrational. He has tried to reason with Mary, arguing that he and many others cross on the stairs all the time, with no negative consequences befalling them. He has tried explaining the origins of the superstition, in order to further argue there is a rational reason for why the superstition came about

[78] (Baron, The Mens Rea and Moral Status of Manipulation, 2014, p. 118)

which no longer justifies the practise. Mary does not accept his arguments and continues to believe that crossing on the stairs is bad luck.

Does Mary's irrational position justify even limited manipulative action John may take towards Mary? I think intuitively it does not. Mary's superstition does not affect John, and Mary presumably has the right to believe whatever she wants. It may not be rational for her to believe in superstitions, but this does not give John the right to manipulate her just because he cannot convince her using arguments or non-rational persuasion. What is missing from this example, in comparison to the arrogant supervisor, is that auxiliary goal which motivates the influence. If, for example, Mary regularly prevented John from performing his job because she refused to walk up the stairs. Or if her belief in the superstition caused her a great deal of distress, we may be willing to consider whether manipulation is justified. But this is the same moral calculus necessary when we recognise that an act will wrong somebody, not when the act does not need any moral justification.

My argument then is that people like the arrogant supervisor may be acting irrationally, but they are not acting irrationally in a way which justifies manipulative action against them on that basis alone. Rather, manipulation may or may not be justified relative to the goal of the influence, and whether or not reasons for influencing them would outweigh the fact that the agent will be wronged. With this established, we can turn to the second sort of case, which are the types of persons Strawson's examples mostly refer to.

Strawson's examples are agents who have severely diminished capacities in the context, not those who are not using their rational capacities in the correct manner. A person in the depths of a mental breakdown is certainly not in the right mental state to be persuaded, if their breakdown is such that they cannot engage with the influencer. Nor would a rowdy, drunk friend we are trying to get home, who is chemically unable to consider arguments. The arrogant boss is not in such a state. He can engage with arguments; he is just so full of himself that he always dismisses them due in favour of his own. The people Strawson give as examples are people who we do not consider morally or rationally responsible in their current circumstances for their own actions. We therefore treat them as persons who act poorly because of physical circumstances and determined causes. They are treated as subjects of influence alone, rather than agents, because their ability to act as agents is compromised. This is not the case for the boss he is still a moral agent, only one with standards we despise.

Can we manipulate these people without facing any case to answer, as it were, on a moral basis? Once again, I don't think so. I argue that either once again manipulating such persons requires good reason, and thus justification, or they are an agent who cannot be manipulated on the trickery account at all.

I would refer back to the example of Casanova. Casanova is a man who does not subscribe to the cognitive ideal that it is better to believe that which is true rather than what is false. Instead, he has a strange set of ideals. He wants to believe only that which will make him more appealing to his current romantic target. Yet, I argued that if Casanova were tricked into the purchase of a car by being lied to about its quality, he would agree he was manipulated. The ideals are objective, in the sense that they are objective standards for good reasoning. Casanova is an unreasonable agent, not just because he believes falsehoods, but because he has no interest in believing the truth at all.

To attempt to act manipulatively, and thus wrong a person, is to act intentionally upon an agent without the intention that the agent does not fall short of one or more ideal which govern their belief, desires or emotions as a direct consequence of the influence. By forming such intent, the

manipulator must have a belief about what the ideal beliefs, desires and emotions are for the agent. Usually, this is objective, and the manipulator hold the target to the same standards as they do themselves, but not always. Therefore, with those who are entirely unreasonable, in that they lack or have severely diminished rational capacity, techniques we may think of as manipulative may not be manipulative at all. To a person who no longer has any interest in believing the truth, who is not merely mistaken but pathologically delusional, deception does not debase them. An agent who knows who they are dealing with cannot have the intent to make a person less ideal by the standards they believe govern the other person. That person has such strange ideals that manipulation cannot take place in the way defined as trickery. Rather they are simply being influenced. An agent who does not know is still acting manipulatively, as they think they are making another less ideal, and thus degrading them from their perspective. You cannot truly trick somebody who is not playing by the same rules of reason as you are. You can only believe that you are, if you are unaware of how they think, and this still says something about your moral character. Therefore, in this context I do not think Casanova or those like him, with the sort of limitations Strawson gives, are able to be manipulated by trickery at all, unless it is on their own terms.

Consequently, I would argue that while many of Baron's observations are correct when it comes to people's usage of the objective viewpoint, manipulation as defined by the trickery account cannot be innocuous no matter what the capacity or 'unreasonableness' if the agent, no matter how defined. Either they are of the correct sort of rational capacity for it to be a wrong, no matter how easily justified, to manipulate them, or they lack the capacity, and cannot be manipulated by a knowing agent. I should also point out that even those who are in such a state that they are reduced agents, may only be like this in certain contexts. A person with a severe mental impairment may need to be treated as a compromised agent one hour, then the next have regained many of their faculties. Children also can be very much realised agents deserving of every respect in one context, but lesser in the next. There are few cases where we can look at a person in their totality and decide they are now deserving of manipulative techniques.

In summary then, Baron provides two arguments in favour of the idea that manipulation can be innocuous. These arguments could apply to the trickery account as I've defined it. The first argument is a set of examples which intuitively are innocuous examples of manipulative action. However, on the trickery account these are likely not to be defined as manipulative acts when carried out with the sort of intent typically underpinning such actions, or are examples I have set aside for further consideration of whether they are manipulative or not, and whether an intuition that they are innocuous is justified in such case. The second argument applies Strawson's objective attitude concept to manipulation. I have argued that agents being unable to be influenced by persuasion or non-rational influence does not justify then manipulating them. Rather, the manipulative action must be justified by a reason which entails the act becomes permissible, albeit still an act which wrongs the person influenced. In cases where agents are compromised, whether by an impairment or by not being a fully developed agent, such as a child, either the agent attempting to act manipulatively knows this, or does not. If they do not, they still have the correct intent and act manipulatively, though they fail to do so. Therefore, they still wrong a person in all important ways when it concerns their choice to act manipulatively, rather than influence using another method. If they do, then they cannot manipulate, as they will not be able to form the correct intent. Therefore, I conclude that both of Baron's arguments can be resolved, and that we should correctly take the approach that manipulation as defined by my amended version of Noggle's trickery account is not innocuous.

Manipulation and Small Wrongs

We should now consider another argument in favour of manipulation being, in some circumstances, innocuous. This argument, stated in brief, is that some acts of manipulation are so slight and insignificant that they cannot be considered degradations or acts of disrespect to the target's agency. The argument is that the wrong I have identified as being always present when a person acts manipulatively is not always present. Therefore, in some cases manipulation would be not only justified in the sense that our moral reasons against acting manipulatively are outweighed, but those reasons would be resolved and disappear entirely. In those contexts, it would be innocuous to manipulate.

One might argue that we do not usually treat people who deceive us as if they have mounted an attack against our dignity, and our capacity to function properly as moral and rational agents. Small instances of deception are present and either accepted or tolerated within a variety of contexts. This includes our personal relationships, both intimate and between strangers.

If a person manipulated me such that I handed to them my life savings, on the promise that they would repay them, and it turned out they had lied to me about their entire identity as a way to swindle me, I may feel the sort of attack on my personal dignity described by Noggle and further elucidated upon by myself. On the other hand, do I feel the same way when a friend tells me there is no beer left in the fridge, and I find out they lied to me so they could snag the final beer for themselves? Intuitively, these are at the very least at drastically different scales.

However, I do not think such examples go much further than that. The wrong of manipulation is that it is an act of disrespect. There are great examples of sweeping disrespect and examples of very slight disrespect. Nevertheless, both are examples of disrespect. We could compare it to the disrespect of not taking off your shoes in a person's home, and the disrespect of stealing from them. Both show a disrespect for a person's living space and the fact that they have given you hospitality, but they exist on a scale. One is more disrespectful than the other.

Where exactly to place each example of manipulation is not a project I want to embark on. Some acts of manipulation are clearly more severe than others, and thus more degrading. Similarly, I do not want to embark on the project of setting out exactly when manipulation may be justified. It is more interesting at this point to consider whether an act can be manipulative but the harm exist on such as small scale as to mean it is no longer sufficient to count as wronging the agent manipulated at all.

I should note that this argument arises in no small part due to the fact that deception is defined as manipulative under the trickery account. Therefore, we inherit a large amount of historic dialogue regarding whether deception is, in all instances wrong. The majority of the examples in this subsection then will regard deception. However, similar logic can apply to other sorts of manipulation, such as temptation and emotional manipulation. Some, like conditioning, we may be more tempted to say are always degrading on the same scale, no matter what external context the conditioning takes place in.

Let's consider a common example discussed in the ethics of deception, the surprise birthday party, stated in my own style for ease of reference and taken from no body of work in particular.

Surprise Party: Kate and Marvin are friends. Kate's birthday is coming up soon, and Marvin has arranged a surprise birthday party for her. Marvin lies to Kate and tells her that he needs her to pick up a parcel for him from an office. Kate goes to pick up the parcel, but instead finds Marvin, Kate's other friends and her family, waiting for her with the office decked out for a birthday party.

Marvin lied to Kate about the reasons he needed her to go to the office. Therefore, he manipulated her. On the trickery account, this means Marvin has wronged Kate. This may be at odds with our intuitions. Assume that Kate enjoyed the party, and especially the fact that it was a surprise party. Kate did not want to know about the party. Lying to her is therefore essential to her enjoyment of the party. We could imagine circumstances where active deception is not required. Perhaps there really was a parcel that she needed to pick up, and Marvin planned the party knowing that she would attend at that time regardless. However, it would be equally unintuitive to state that Marvin has a moral duty to find such circumstances and exploit them, rather than lie to create circumstances where he can surprise Kate. The intuition is that lying in this case is not just justified but has nothing morally suspect about it at all. Of course, we could imagine a situation where the surprise is not received positively. Personally, if I were to have a surprise party thrown for me, I would not enjoy it at all. Kate could be the same. However, this would be to focus on the external elements of the context, not the moral characteristics of the act of lying itself.

We could consider Kate as having given tacit consent based on cultural norms here, which would transform the act of lying into a moral one. Just as consent to a boxing match makes punching a person in the face no longer morally suspect, cultural context could provide tacit consent. Explicit consent could also be given if Marvin and Kate had discussed the idea of surprise parties, or Kate had enjoyed a surprise party thrown for her in the past. Tacit consent for lying would be found in the cultural construction of the surprise party itself. A "surprise party" is a cultural construct. We recognise that for birthdays or similar culturally significant events, the joy of friends and family coming together without your input to celebrate you, or your achievements is something joyful and desirable. Consequently, unless consent is withdrawn, it can be assumed that you will *want* to be lied to in order to bring this about.

We can consider another case where lying is culturally accepted, at least in the culture I am immersed in, I cannot speak for other cultures. One would be concealing an engagement ring when an agent plans to propose to their significant other. While people do have open discussions around marriage, there is still the cultural expectation that the ring, and the proposal can be a surprise. In the course of writing, I proposed to my own partner, and in doing so hid the engagement ring, its purchase, and my intentions until the opportune moment for a romantic proposal arose. I did not deceive but would have done if my partner had almost come across its hiding spot or asked me outright if I intended to propose. I do not see myself as being a manipulator because of my actions.

If we accept that these cases are cases of deception which appear to be innocuous, we must either accept that these are cases of manipulation, and that manipulation can at times be innocuous, or we must see these as cases which have been transformed by consent such that they are no longer cases of manipulation at all. The latter approach I prefer, but it needs further explanation.

I argue that we must remember that the reason the cognitive ideals are defined as ideal in the first place is due to their practical value in allowing agents to fulfil their goals as rational agents, as well as realising them as rational agents. Where tacit or explicit consent exists for deception, or other elements of manipulation, the influence does not degrade the target, even though it may instil a non-ideal belief, desire or emotion. Rather it was the choice, or assumed choice, of the target of the influence, to allow non-ideal beliefs, desires or emotions to fulfil a goal of that agent which could not otherwise have been attained.

There is a counter-argument to this idea, that it makes influence non-manipulative if it is explicitly desired by the target. We may still want to call such influence manipulative, and still call it morally suspect. For example, imagine that a woman's daughter runs away from home, and the woman

desperately wants to believe that her daughter is alive, happy and safe wherever she is. We might know that her daughter sadly died after leading a miserable life on the streets when she ran away. We are then faced with a dilemma of allowing the woman to live in blissful ignorance or delivering to her a harsh truth she does not want us to reveal. This treads on theory of self-deception, and whether there is a moral obligation to seek and believe the truth, regardless of how unpleasant those truths may be. I do not think our argument has to go so far as to require examination of these issues. One element of the surprise party example and the engagement example is that we do not expect the deception to be permanent. The deception serves a very limited purpose which does not cause harm to the foundation of our ideals, that we want to believe only that which is true, desire that we have reason to desire, and feel emotions that are appropriate. Rather it is a temporary exception which will end in revelation.

We have then, a set of conditions which, if met, results in influence which makes us fall short of an ideal not being defined as manipulative. Firstly, there must be explicit or tacit consent for the manipulation. Tacit consent occurs in circumstances where it is culturally acceptable for an agent to be cognitively unideal to serve some other purpose desirable to the agent. Crucially, this tacit consent can be withdrawn. If a person like me explicitly states they do not want to be lied to as a surprise party is not something I would enjoy, then one does not have consent to influence me in that manner, and it would be manipulation if you knew my stance and proceeded to influence me regardless. Secondly, the influence must be limited in scope only to achieve that aim which has been consented to. Thirdly, it must be expected that in achieving that aim, the non-ideal state must be corrected. Revelation must be built in to aim of the non-ideal state in the first place.

When these conditions are met, I argue that examples such as the surprise party example can be seen as non-manipulative influence, despite making the target agent non-ideal for a time. These are limited in scope and provide a reasonable basis for why we intuitively do not see such influence as harmful or immoral, or even standing in need of justification. However, it does not extend to examples of self-deception, or agents who permanently wish to be made non-ideal and thus provide consent for it.

Conclusion

Having considered these examples and arguments, I maintain that we should consider manipulative acts as always having pro-tanto reasons that they should be impermissible, by virtue of them being manipulative acts. Manipulation always involves wronging the target. This wronging may be justified by other moral reasons, such that manipulative acts are permissible in many contexts. Nevertheless, I maintain that there is no such thing as an example of innocuous manipulation under the trickery account.

The Moral Responsibilities of the Manipulated

Let us cast our minds back to the example of the Cult Leader, who employs manipulative techniques upon his followers to get them to serve his every whim. He casts himself as a messiah, the second coming of Christ. Consequently, he must grow his church and any who deny him or otherwise go against his orders, are going against the will of God. He abuses his power over those who are his followers. The Cult Leader in this example is as clear an example of a manipulative agent as one can imagine. However, no self-respecting cult leader works alone to grow their flock. Instead, they send their followers out into the world to gather further members. In this case, they get their followers to employ the same techniques. We know the moral position of the Cult Leader, but what can we say about the followers themselves?

Obviously, their positions are different. Firstly, the followers are themselves the victim of a wrongful act, which has then motivated their actions. Secondly, their internal mental position regarding whether they intend to make a person more or less ideal will be different. The follower, if we presume they have been manipulated as well, has been indoctrinated into the cult's worldview. They are sincere, whereas at least in this example, the Cult Leader is not. This does not entail that the followers, if they employ inherently manipulative techniques such as love bombing, do not have moral responsibility for those actions, or do not act manipulatively. That remains an open question. However, in regard especially to the ideals for belief, they cannot manipulate in the same way. If the cult-leader lies when they convince their followers of a belief *p*, as they know that *p* is false, the followers do not do the same if they convince others of *p*, as they sincerely believe that *p* is true. In fact, we might think of the manipulator as baring greater responsibility the more people believe falsely in *p*. We might be inclined to think that the cult-leader has not just manipulated one person, they have manipulated many more by proxy.

Another more grounded example might be that of a rumour. Imagine you maliciously lie about a rival colleague in your workplace, perhaps claiming that you saw them cheating on their spouse. If this spreads around the office, then we could say you are causally responsible for the beliefs of all of those who believe the rumour as a result of your influence, even though you only directly influenced a single person. You are the source of the falsehood, and thus bare responsibility for the epistemic misstep of all your colleagues as the rumour spreads. You may also be culpable for it. Yet while you may be responsible as the source of the rumour, this does not mean it is true that you have manipulated the entire office. The barrier to this conclusion is that on the manipulation as trickery account, the influence must be intentional. Perhaps you only wanted the first person you told, your manager, to think worse of your rival. You did not intend for the entire office to believe your lie. In fact, you never directly influenced them at all. The influence was indirect.

Is this an acceptable consequence of the account? I believe it is. We can find moral wrongs in the spread of misinformation, especially on a mass scale, which does not require the specific moral wrong of manipulation to explain and satisfy our intuitive attribution of wrong to people who spread misinformation. If a person only intends to deceive one agent, we cannot claim that they intentionally influenced others the deceived target then spreads the false belief to. This is not to say that a manipulator cannot act to intentionally manipulate a larger group of people through the use or others. For example, if I tell a group of friends a false rumour with the intention that they will then go on to influence others to believe it. In this example I would be manipulating a larger group of people in the same manner that I would if I had stood in the office with a megaphone, influencing intentionally all within earshot. This will be relevant when we move on to consider mass-manipulation in the next chapter.

For now, we can summarise. Those who are manipulated by way of deception who then spread the false belief, are not acting manipulatively if they genuinely think the belief is true. The manipulator, unless they intend those they deceive to spread the false belief to others, does not manipulate anyone other than their intended target. The cult leader then, depending on their instruction, does manipulate through their followers if those followers are instructed by the cult leader to recruit new members. However, the cult members may lack the intent to manipulate. They do not manipulate personally but are instruments of the cult leader.[79]

[79] We should remember however, that this analysis would not apply to techniques which are necessarily manipulative as they cannot be carried out properly without manipulative intent. I argue love bombing is one

Another interesting question is how morally responsible a person would be for acting based on manipulative influence. It is intuitive to think that a person at least has less responsibility for their actions if a significant motivation for that act, was the consequence of manipulative influence. After all, the person being manipulated is a victim. Consider the following example.

Cult-motivated Theft: Cindy has been attending regular meetings of a church known as "The New Baptist Revolution". The church is led by Robert, who fashions himself as the second coming of Christ. Part of Robert's teachings is that the church needs funds, as well as to retreat to a secluded place away from the sinfulness of the world. The church teaches that non-believers are evil and will go to hell. Thus, it is not morally wrong to steal from them or hurt them. Consequently, Cindy steals a great deal of money from her parents before absconding to live in the new compound with her fellow cult-members.

The moral wrong of manipulation as trickery is the disrespect and attack to the dignity of the targeted agent by prevent them from operating as an optimum agent should. With this in mind, we can see why we might evaluate Cindy as less morally culpable, that is less worthy of blame, than if she had not been subject to manipulation. We can refer back to Strawson's 'Freedom and Resentment', and that category of people who have diminished agency, be they children, the severely mentally afflicted, etc. We take the objective attitude towards these people, viewing them not as agents to be reacted to and evaluated personally, but as forces in the world without their own will. As Strawson states, to view a person wholly with the objective attitude means "then though you may fight him, you cannot quarrel with him, and though you may talk to him, even negotiate with him, you cannot reason with him. You can at most pretend to quarrel, or to reason, with him."[80] If we take this attitude towards Cindy, we view her as an agent so compromised by manipulation that we can no longer view her as a moral agent at all. She is like a storm, or other natural phenomena, a person who is more like a detrimental force in the world than an agent. I compared manipulation to temporary lobotomy, well to extend the metaphor Cindy is wholly lobotomised by the manipulator, and so should be considered as an impaired agent.

This analysis is extreme. We are assuming here Cindy is manipulated in a manner so severe as to affect her entire being and change her into something entirely subject to another person's will. This is the sort of person who, in common parlance, would need to be 'deprogrammed', because they have been 'brainwashed'. Most manipulation will not have such consequences. Consequently, we'll also think of them as more morally culpable.

Consider the following example:

Toxic Relationship: Bob and Mandy are friends. Bob forms a new romantic relationship with a woman named Polly. Polly dislikes Mandy and is jealous of the friendship between Bob and Mandy. Polly starts to lie to Bob, claiming that Mandy is hostile to her when Bob isn't present. She insinuates that Mandy has romantic feelings towards Bob and is jealous of his and Polly's relationship. She uses this claim to make Bob feel guilty for maintaining his friendship with Mandy. As a consequence, Bob starts to distance himself from Mandy, without Mandy knowing why, which causes her distress. Bob starts to break promises to Mandy, and eventually ends the friendship between them.

---

of these techniques, but there are more. This also does not entail that a cult member could not lie consciously in the service of the cult for the purposes of recruitment, if they have justified that lie to themselves. Rather, the analysis above regards whether the sincere spread of what was an intentional falsehood entails the followers act manipulatively.

[80] (Strawson, 2008, p. 10)

This example is similar to **Cult-motivated theft**, in that a person wrongs another directly because they have been manipulated. I take it that the ways in which Polly influences Bob to end his friendship with Mandy are at least partially manipulative acts, though of course in such a scenario, a mixture of all different types of influence would likely be used. I also hold that the actions Bob takes towards Mandy, such as breaking promises to her, and treating her differently than their friendship would mandate, are actions Bob takes which count as wronging Mandy. Bob is not a compromised moral agent as a result of being manipulated in this case. At the very least, we could understand if Mandy, even if she knows that Bob has been manipulated by Polly, still believes she has been wronged by Bob.

Mandy could take the objective attitude towards Bob, seeing him as nothing more in this context than a puppet for Polly's desires. However, the objective attitude is not entirely justified in this case. It is reasonable for Mandy to have expectations that Bob would value their friendship, and have moral and rational principles sufficient for him to respond to Polly's influence in a way which respects these obligations.

Similarly, in cases of domestic abuse, the victim of the abuse may behave in morally repugnant ways towards others as a result of the abuse. This may include manipulative acts. Bob seems more responsible and Cindy from **Cult-motivated theft** seems less so. I suggest that the reason this intuitively seems the case is that Cindy's agency is more compromised than Bob.

There is intense similarity between cases where a person is subject to manipulative influence, and cases of addiction, compulsion, or other habitual behaviours. We may not consider a person morally responsible for actions taken under intense compulsion if it is sufficiently intense by some unknown calculation. However, as a rule we do recognise that persons under less intense influence remain morally responsible, balancing the influence of other forces with their own reasons desires and emotions. The thinking here is similar to that proposed by Susser et al in '*Online Manipulation: Hidden Influences in a Digital World*'. It is true that those who are subject to manipulative influence retain their 'freedom of choice', but face pressures and interference in their thinking which hampers to extent to which they can firstly, make those practical decisions, and secondly, impacts the extent to which they are the origin of those decisions. While I reject their assertion that phenomena such as temptation cannot be defined as manipulative, we can use their work when considering the moral responsibility of those who are manipulated.

While the manipulated remain morally responsible for their actions, this is on a sliding scale. When a person has been entirely captured by manipulative influence, we can hardly treat them as an agent at all. Rather, they are treated as an extension of the manipulator's agency. At the other end of the scale, usually regarding information manipulation, a person appears to retain their agency almost entirely. Decisions made based on false information must be evaluated morally as if the information were true.

In '*Information Manipulation and Moral Responsibility*', Todd Long argues that manipulation which only alters the information an agent uses to make a decision does not affect the degree to which they are morally responsible, but their moral culpability.[81] To paraphrase his argument drastically, well-reasoned decisions we would usually ascribe moral responsibility to can be made on sets of true and false information. Whether the information is true or false does not change the fact that:

"In both cases, you deliberate on the basis of information that you have non-culpably acquired and which you have excellent reason to believe is true. In both cases, your action is consonant with the

[81] (Long, 2014, p. 162)

desires, preferences, and values that make up your character. In both cases, the way you deliberate to your decision is the same. Thus, your way of deliberating and acting on the basis of that deliberation is not manipulated at all. All that is manipulated is the information that you have to go on in your deliberation."[82]

Manipulation is trickery. Long argues that as long as the trickery doesn't interfere with your ability to make a decision, so long as it only changes your beliefs, whatever your decision is, it is one which has all the features necessary for you to be morally responsible for it. If then, you are told a lie about a friend, and treat that friend poorly as a result, you remain morally responsible for that poor treatment. However, you may be less morally culpable, that is worthy or either blame or praise for the action, as it was based upon a false belief it was nevertheless reasonable for you to hold. However, this is not always the case.

This accords with what I believe is sensible intuitions regarding culpability and responsibility. That a person is the victim of manipulation clearly does not, in all cases, rid them of any moral responsibility. For example, consider if I knew a person was very homophobic. For whatever reason, I manipulate them into believing falsely that one of their close friends is a gay person. The homophobe wrongs their friend as a result of my manipulation. Clearly, the fact that the homophobe was manipulated in this case, and performed what we can safely assume is a morally wrong act, is not sufficient to excuse them of any moral responsibility or moral culpability. Whether they were manipulated or not, their act was wrong. However, we can imagine cases where a person is manipulated and as a result is less culpable of wrongdoing or praise for their actions. From this example we know two principles which can excuse a manipulated agent from morally wrong acts they might perform as a result of being manipulated.

1. The manipulation is sufficiently detrimental to the ability of the manipulated agent to make decisions which are consonant with the desires, preferences, and values that make up the agent's character. Such that we can consider them to be an impaired agent.
2. The action carried out by the manipulated person would not independently be morally wrong, regardless of whether or not the manipulated person was manipulated.

If either of these are true, the agent is at least less morally culpable. Cindy then, if we take her to be entirely brainwashed by the cult, is an impaired agent. We can also evaluate Bob's conduct. A friend who wrongs another friend based on lies they've been told about that friend only acts wrongly if it would have been wrong regardless of the truth or falsity of the information. I would argue Bob does wrong Mandy through how he acts, regardless of the truth or falsity of Polly's claims.

However, there is one more element to consider in the latter example, what if the lie the friend believed was one which was outrageous, such that it was unreasonable. Long's examples specifically mention that in each case you have 'excellent reason' to believe what you believe. We should consider then whether a person who is manipulated can be morally culpable for falling for the trick in the first place. Let's consider another example.

Banking Scam: Merriam works for a large bank. She receives a call from a woman, Kelly, who claims to be escaping from an abusive relationship. Kelly claims she needs money urgently transferring from her existing account into a new one with a different bank. She claims to be unable to answer her more advanced security questions, such as a passcode sent to her phone, because her abusive husband has the phone, and she is scared to try to retrieve it. Merriam has been trained in the bank's security procedures and knows that she should not transfer funds to another account without

[82] (Long, 2014, p. 153)

the customer passing the correct security procedures. However, Kelly insists that not doing so will leave her destitute and in life threatening danger. With the claims of urgency, and with the story she has been told, Merriam feels pressured to transfer the funds, and does so. It turns out 'Kelly' was a scam-artist, and had stolen the identity of the real Kelly, who has now lost a substantial amount of money. Merriam loses her job because she transferred the funds.

Merriam in this example is a victim of manipulation. She was manipulated in regard to her beliefs, as 'Kelly' repeatedly lied to her, as well as in regard to her emotions, as her attention was taken from what should have been salient matters, the fact that 'Kelly' couldn't provide the proper security measures, to the manufactured urgency of the situation. However, Merriam is held responsible for her decision by the bank. Does she deserve blame for the loss of funds to the real 'Kelly'? Intuition would say that she does. She should have been more sceptical and stuck closely to the banks security procedures. We could claim that Merriam fell short of an epistemic ideal, and instead fell into the vice of gullibility.

Merriam was definitely a free agent. She retained enough of her capacities that she should have recognised the manipulation for what it was. Part of being a rational agent is having responsibility for evaluating the influence you are subjected to. A person who does not evaluate with any care information they are told, is a poor agent, responsible for the consequences acting on that poor information entails. Yet I would also say that because one of the manipulative techniques used was emotional, not based on belief, Merriam's culpability is diminished somewhat, though not entirely expunged. She was somewhat compromised in her ability to act as a rational agent as manipulation of this sort can be very powerful. That should not be understated.

The contrary view to this position would be that Merriam isn't culpable at all. After all, scams and fraud can trick people from all manner of backgrounds. One might argue that manipulative techniques are effective against all agents, not just gullible agents, though some may be more susceptible to manipulative influence than others. A real-life example could be the Ponzi scheme of Bernie Madoff. Madoff was an extremely respected Wall Street investor who was known giving large returns on investments. Banks and other wealthy institutions were exposed to immense loss of investments when, in the 2008 financial crash, Madoff's scheme collapsed. Rather than being the market genius he purported to be, Madoff paid new investors large returns by using the money of new investors. Hence, the entire investment operation was a pyramid scheme, where Madoff required increasingly large investments to continue to pay his old investors. We might think that surely, the various individuals involved in trusting Madoff, experts in their financial fields, should have known at the time that Madoff's large, consistent returns on investments, regardless of how the market was actually performing, were suspicious. [83]

What we are looking for here is what it means to have failed due diligence of evaluating influence an agent is being subjected to. At what point is accepting absurd influence a failing which means one is culpable for immoral acts performed as a result?

Alexander Greenburg considers what it means to be responsible for one's beliefs by comparing it to the definitions of criminal negligence in his paper '*Epistemic Responsibility and Criminal Negligence*'.[84] Greenburg argues that the most promising traditional account for how we can be responsible for our beliefs is that we have a direct control, though not a voluntary one, over our beliefs. This neatly applies to "reason-responsive" beliefs which are subject to a process of rational

[83] (Treanor, 2008)
[84] (Greenberg, 2020)

evaluation. A belief is "reasons-responsive just in case it's formed or revised by a reasons-responsive process".[85] For our understanding, we should consider a reasons-responsive belief to be one which follows from the capacities of an agent to evaluate that reason. However, this would exempt delusional beliefs which do not arise from our capacities as a rational agent but instead from some non-reason responsive process. This is intuitively appealing if we consider beliefs which arise from causes like mental illness. Greenburg specifically notes that this would also exclude beliefs which arise from manipulative techniques, such as state indoctrination or gaslighting. Yet there are some beliefs we have which arise from non-reason responsive processes that Greenburg argues we should want agents to be culpable for, such as bias and wishful thinking. For example, a person with biases based on racist or xenophobic beliefs. The typical view that a belief being 'reason-responsive' means we are responsible for it, and we cannot be responsible for non-reason responsive beliefs may then be flaws.

Greenburg proposes his own account by developing an account of criminal negligence by H.L.A Hart. Relevant for our discussion is that he answers an objection to Hart, that Hart's account does not adequately explain why failures of reasoning by defendants accused of criminal negligence speak badly of them, by arguing that "The negligent defendant's failure to exercise his capacity to recognise risk manifests insufficient concern for others' interests."[86] Those guilty of criminal negligence are not just able to recognise the risks of their conduct in a position of care over something, and fail to exercise that capacity (failing to utilise their rational capacities in the relevant sense), but do not do so because they have in inadequate regard for manging the risk to others interests in the first place. This insufficient concern manifests in their actions. An example given is that forgetting you had agreed to meet a person for coffee several times, reveals a general lack of regard for meeting the person in the first place. If you truly cared, you would remember.

The account for criminal negligence, when a person is responsible for a lack of care as to the consequences of their actions, can then be applied cleverly by Greenburg for what we are epistemically responsible for the consequences of our beliefs.

This results in the following test:

"Justified: S is epistemically responsible for a justified belief only if the fact that the belief is justified manifests S's sufficient concern for the truth.

Unjustified: S is epistemically responsible for an unjustified belief only if the fact that the belief is unjustified manifests S's insufficient concern for the truth."[87]

Greenburg then applies the test to examples of gaslighting (manipulation under the tricker account) and confirmation bias. In the bias case, the believers "overriding concern is one's interest in one's initial judgement being correct", not a concern for believing the truth. Therefore, the believer in this case is culpable for their belief. On the other hand, the victim of gaslighting *is* concerned with believing the truth. The unjustified belief a victim of gaslighting may hold is caused by the manipulator's influence abusing her desire to believe the truth, but providing her with faulty information. The same is true with propaganda. If a person sincerely believes the government's position on, for example, a war, then whilst the justificatory reasons for the belief may be mistaken, the believer does have a sufficient concern for the truth to exculpate them.

[85] (Greenberg, 2020, p. 95)
[86] (Greenberg, 2020, p. 105)
[87] (Greenberg, 2020, p. 107)

Those who would be targeted for criticism as a result of this account would be those who in believing an belief *x*, do not show through their actions a genuine concern for the truth of *x*. Greenburg gives the example of a conspiracy theorist who appears through their beliefs in varying conspiracies less concerned with the truth of the theories than the idea of secret plots existing in the first place, or the preoccupation by wanting to be part a select, enlightened group who know what's 'really' going on.

What for Greenburg is a clever sequence of steps which results in a way to judge whether a person is epistemically responsible for a belief they hold can aptly be applied to answer our question regarding the culpability of those who have been manipulated, such as Merriam the bank teller.

When a person is successfully manipulated in regard to their beliefs, they are either made to believe something that is false or made to think a belief is more or less relevant to the context than it actually is. For a person to be exempt from responsibility for these beliefs regarding truth or relevancy, they must be shown to have sufficient regard for whether the belief is true or relevant.

For emotion, a person is successfully manipulated when they are made to feel an emotion which is inappropriate in the given context, or when they are made to feel an emotion which makes more salient aspects of the context which ought to be less or more salient. To be exempt from responsibility, they must have shown due concern for whether their emotions are appropriate or whether they make salient that which they should.

For desire, a person is successfully manipulated when they are made to desire that which they have no rational reason to desire, or are made to desire strongly that which they have reason not to desire. To be exempt from responsibility, they must have shown sufficient concern for whether their desires are rationally supported.

In each case, we can consider the manipulated at fault if, in being made to fall short of an ideal, they do so at least in part due to an insufficient concern for the ideal in the first place. In practise, what is sufficient concern is an introspective examination of the products of the influence.

In the Madoff example, the financial experts who were manipulated by Madoff may have responsibility for the losses caused. They were tricked, yes, but we could accuse them of not sufficiently interrogating their own beliefs. Madoff made gains above and beyond what could be expected of the markets. Those with less financial expertise could be forgiven as having examined their beliefs with the resources available to them, but we can accuse at least some financial experts of a wilful ignorance. They were happy with the returns they received and did not wish to further interrogate how they got them. Therefore, although they were manipulated, many will be culpable for any immoral actions they performed as a result.

In regard to **Banking Scam**, Merriam was manipulated by 'Kelly' into believing falsehoods, as well as feeling an emotion which made less salient features of the interaction, that 'Kelly' could not provide the required credentials to make a transaction over the phone. To evaluate Merriam, we should consider the context she is in. Did she provide sufficient interrogation of 'Kelly's' lies, or did she believe them at face value? Did she introspect and wonder why 'Kelly' was invoking a feeling or urgency and pity, and 'Kelly's' reasons for doing so? Presuming that Merriam had been trained by the bank on the tactics of such scams, and the reasons behind security protocols, Merriam would be guilty of not sufficiently interrogating such reasons and not taking her responsibilities as a rational agent in this context seriously. On the other hand, if we alter the example such that 'Kelly' *could* provide the right credentials –perhaps the scammer had also stolen the real Kelly's phone to provide

the correct code– then we might consider Merriam as having sufficiently examined the influence she has been subjected to.

Finally, let us consider Bob in **Toxic Relationship**, who has started a new romantic relationship with Polly, who despises Bob's friend Mandy, and as such manipulated Bob to feel guilt for his continued friendship with Mandy. Bob is responsible to the extent that they do not rationally examine whether it is appropriate for them to feel guilty. If they do, and mistakenly believe that yes, guilt is appropriate, then they are not responsible. However, if they do not sufficiently examine the appropriateness of the guilty, perhaps out of a vested interest in continuing to be with their new partner, then they can be held responsible. In regard to Bob's beliefs, if Bob does not sufficiently consider whatever evidence Polly provides to prove Mandy has romantic feelings towards him, then we can hold him culpable for believing Polly's lies.

Of the manipulator in all of this, we can add a new condemnation. The manipulator is not just responsible for their own demeaning mistreatment of those they manipulate, but for causing the manipulated to fail in their epistemic duties, and all the practical consequences that may entail.

To summarise, we have considered two questions. The first is whether people who are manipulated, manipulate in turn and are morally responsible for that manipulation. An example would be the spread of misinformation. I concluded that those who appear to manipulate as a consequence of being manipulated, lack the correct intentions for their actions to count as manipulation. There are other epistemic wrongs we can accused them of, such that intuitional condemnation of a person who ignorantly spreads misinformation can be accounted for without the need to also accuse them of manipulation. The general exception to this is the use of techniques such as love-bombing, which have an essential intent which is manipulative.

The second question is whether the manipulated are morally responsible for other actions they perform as a consequence of the manipulative influence, especially if being successfully manipulated was due to a failing of their own. Considering manipulation on the basis of belief, or information manipulation, we can easily find examples where being manipulated does not affect a person's moral responsibility for their actions. Considering other sorts of manipulation, as well as Strawson's objective attitude, we can also identify examples where a person has been so diminished as an agent, we cannot ascribe moral responsibility to them. When taking the objective attitude is not appropriate, we have established that an agent is not culpable for actions taken due to manipulative influence, only if they can be shown to have due regard for the ideal corresponding to the type of influence they faced. The victim of manipulation as trickery, is only not morally responsible for actions taken because of the trick if they showed some resistance to the influence, examining it with the required scepticism agents should utilise when encountering any influence. A person who does not show due diligence is culpable, though the fact that they were manipulated should certainly be a mitigating feature when we evaluate them. What exactly "due diligence" means for a given contest will obviously vary.

Summary

In this chapter we've considered the key questions regarding the ethics of the manipulation as trickery account. We considered the moral status of manipulation, and I have argued that manipulative acts are always a wrong committed against the target. Whilst manipulation is often justified, I don't believe this wrong is ever resolved, to the extent where it is not still a moral wrong committed, even if the manipulative act also does a world of good.

We've also considered more tangential, but still extremely interesting questions which follow from the trickery account. We've considered whether agents who act wrongly as a result of manipulative influence have a diminished responsibility, or diminished culpability as a result. These are key questions with real applicability. I have shown then that the definition of manipulation as trickery provides no barrier to sensible conclusions about its ethics. In fact, I would argue the trickery definition invites an ethical account which satisfies our existing intuitions about manipulation whilst simultaneously allowing room for sophisticated and specific explanations regarding more niche examples.

Section Two

Chapter Four: Mass Manipulation

Introduction:

At great length, we have examined the account of manipulation as trickery. Until now the majority of the examples we have used have been examples of manipulation on a small scale. Either they are between the manipulator and a single target, or perhaps multiple targets who are still subject to the influence in a more intimate context. For example, a car salesperson and the couple to whom he is selling a car. I have defined this as agent-to-agent manipulation.

It should not be controversial to say that an account of manipulation should also apply to less intimate contexts. When it comes to the real implications of manipulation, society at large is now much more concerned with manipulation which applies to vast swathes of people. We are concerned with the manipulative influence peddled by public figures such as ideologues, politicians, preachers, and extremists. While we live in interesting political times, the existence of such figures and our concern over their methods is not new. Rather, concerns about ideology are centuries old. However, new communications technology has given more people access to large audiences. Radio and television were such technologies, but they were comparatively easy to regulate and moderate. Even when governments declined to regulate them, access to audiences through these means was somewhat limited by the resources and connections they required. Now, any person with a smartphone can reach an audience of millions, and while many have tried, it is difficult to fully regular the open internet.

Beyond our worries about the actions of individual agents, and their influence over groups of millions of people, we also worry about the platforms those influencers operate on. We might identify one individual on social media as a manipulative agent and rightfully be concerned by their influence. Yet we might be more concerned with the fact that the platform that single agent recommended us that content, and whether those who designed it had ulterior motives. That topic is for the next chapter, where I will consider the problem of manipulation through design. For now, this chapter will focus on how the manipulation of large audiences works, and how it may differ from more intimate contexts.

On first glance, we might not be concerned about 'scaling up' the account. Consider manipulation by deception. Telling a lie to a single person is at first no different to lying to a crowd. I need only raise my voice. Just as well, if I had a powerful enough megaphone, I could tell a lie to everybody on the globe with the capacity to hear and understand my words. The tools of mass media are just this hypothetical megaphone, a form of amplification. One approach then would be to say that mass-manipulation is just agent-to-agent manipulation done many times simultaneously, and need not be analysed as anything more complicated than that.

There is one example I have used multiple times which is of mass-manipulation. I will reiterate it formally here for ease of reference:

Apocalyptic Preacher – A television minister preaches that the end times will soon be upon us. On the same show, she advertises various emergency food rations, shelters and weaponry. She warns her flock that the prophesised end of days could happen any day, and reminds them of the current

offers on equipment to survive it. She has been doing this for the last ten years, no end of days has yet come. She is quite wealthy.

What differences are there which might resist a straightforward application of the trickery account to this example? Firstly, the preacher does not know much about her audience. She does not know the size of the audience, nor who they are beyond whatever research her production team has provided to her. Let us assume that the preacher does not film her television programme in front of live audience, but instead speaks directly to the camera. Let us also assume she is a very successful preacher, with her show watched by millions from around the world. Her show is broadcast in several countries, and she has recently launched a YouTube channel, where clips of her sermons are accessible to, all things considered, anybody with access to the internet. We know who the preacher, we can alter our example so that she has whatever traits we want her to have. Yet we will struggle to understand an ever-changing audience of a huge number of agents. Thanks to the internet, the 'target' agents of her influence are essentially anybody with the capacity to access and understand her recorded sermons. That is a huge hypothetical audience.

I think it is important to note this is not a novel problem. Medieval historians wrote books filled with politically motivated lies about historical events. They were manipulating just like the preacher, without knowing who will actually *be* influenced by their writings. After all, they likely didn't envisage us, yet they lie to us all the same.

To my mind then, mass-manipulation raises two problems for the trickery account specifically. The first is that in examples such as **Apocalyptic Preacher**, the audience is ill-defined. The challenge from this fact to the trickery account is that the manipulator must have intent towards the target of their influence that the target be led astray, or more specifically lack the intent that they are not led astray as a result of the influence. We have referred to a more general notion of this previously as manipulative intent. From now on, I will use the term specifically to refer to the intent required under the trickery account. If the manipulator knows very little, or nothing at all, about their audience, then we could question the form of this intent, and whether it can exist at all. We can call this the *undefined audience problem*. We will need a picture of what it means to have an intent to an unspecified set of people. We will need to consider whether influencers can have a general intent towards the undefined audience, and to whom that intent may be limited.

The second issue is that some acts of influence are not the product of a single person, but a group of people all who might have different intentions. A practical example would be a politician who gives a speech we may view as manipulative from our position as a receiver of the speech. We might hear the speech and view the politician as acting manipulatively. However, we also know that no modern politician works alone. The politician will have had a team of speechwriters. Those speechwriters will themselves be beholden to officials from the politician's political party, who will have a say in the content of the speech. Those party officials may themselves be in discussion with the party's donors, and so on. Many hands may have worked on the words the politician eventually delivers. It is not a product of one agent, but many, all who may have differing intents as to how they want to influence the people who hear the speech. We might think the speech is manipulative because it contains lies. However, some of those who wrote the speech might genuinely believe its content. Other will not. If we had perfect knowledge of these varying intentions, with some intending to lie and others not to, how would we evaluate the act? We can call this the *undefined manipulator* problem.

In a way, these two issues are not direct challenges to the trickery account, but problems with implementing it. To evaluate whether a manipulative act has occurred, we need to know the manipulator and their intent. The undefined audience problem makes us question what the intent of

a manipulator who directs their intent towards a large audience looks like, and whether it could be coherent. The undefined manipulator problem asks whose intent we should take to define the character of the act of influence, if any at all.

We face two challenges then, mirror images of each other. First, we must consider how to define manipulative intent when applied to unknown, or less detailed audiences of targets. Second, we must consider the intent of the influencer in cases where there may be many agents taking part in an act of influence.

The Undefined Audience Problem

The preacher does not know the exact details of the people she intends to influence. However, she clearly has manipulative intent towards at least a part of her audience. We must presume then, from the offset, that intent can be generalised. For example, if I am walking through town and a person attempts to mug me, threatening me with a knife and asking for me to turn over my valuables, I might shout "Help! I'm being mugged". I'm trying to influence some agents to come to my aid, but I don't know who is able to hear me. I don't even know if anybody can hear me at all. Clearly though, I'm trying to influence, but with a non-specific intent.

The most obvious approach to manipulative intent is to consider manipulators as having such intent towards all the people who might be communicated to by the act of influence. The preacher intends to manipulate everybody who hears her words, whether they be in person or recorded, streamed live on television or the internet, or pre-recorded. This is actually a persuasive reason to hold that the cognitive ideals, as standards to hold rational agents to, should be considered objective. Without specific knowledge of the specific ideals an individual agent holds to, as the preacher knows nothing about them, in order to have manipulative intent she must believe she is leading them astray in a generic way. She assumes the people who watch her sermons are rational agents, and thus intends to make them fall short of the ideals in a generic fashion.

On the face of it this obvious approach functions well. However, there are some caveats. For example, the preacher can have co-conspirators, people who are aware of her manipulative intent but still hear her sermons. She will have a crew who help her produce them, camera operators and the like. If they are 'in on it', so to speak, it would make little sense to say that the preacher has manipulative intent towards them. She knows manipulation would not work on them. If a person knows you are lying to them, and you know that they know for sure with no doubt that you are lying, you cannot make a sincere attempt to deceive them. The agent must believe at least the possibility of success. If we accept the existence of co-conspirators, or those in the know about the manipulation, we can distinguish between the actual audience who are subject to the influence, and the intended audience who the influencer subjects to influence with intent.

To take this further, we could assume that the preacher knows that some of her actual audience, those who will see her videos, will be sceptics, or those of different religions who are watching her sermons for other reasons. Perhaps they are trying to expose her scheme to sell survival items by exploiting the fear of apocalypse. One view would be to treat these agents as alike to the co-conspirators. However, note that in order to make a sincere attempt at manipulation I have said there must be a *possibility* of success. The preacher does not know the specifics of her audience. While she might assume some of her audience will actually fall short, she intends generically. For all she knows, even those sceptics might be able to be manipulated. With some forms of manipulation, such as emotional manipulation, there may be no set of agents who the manipulator knows will be immune to the influence by virtue of their knowledge of its occurrence. As noted in previous

chapters, knowledge of the manipulative influence can with some forms of manipulation only provide partial protection, if any.[88]

Her intent then, can be defined as an intent to manipulate all those who could be manipulated, no matter how remote the possibility. This would exclude those who cannot understand her words, or those who she has knowledge of, such that she knows they know she has manipulative intent towards her broader audience. However, because she has such limited knowledge, her intent can be extremely general to those she has no knowledge of.

The preacher casts a wide net with her mass-manipulation. This is just a consequence of manipulation, or any influence for that matter, in the context of mass manipulation to a wide audience. When I yell for help, I have a general, non-manipulative intent to influence anybody who can hear me and understand what I am trying to communicate. Just so with manipulative influence.

Generic intent exists then. However, going forward, we can understand intent as more or less generic. For example, a politician may give a speech with two different intentions. They may tell a lie in their speech which is intended for their 'base' of supporters, who will believe it emphatically, while also intending for this to be a red herring for any detractors in the audience. They seek to manipulate in two ways. Firstly, by deceiving one section of the audience, secondly by distracting another section of the audience who he hopes will focus their attention on refuting the lie, rather than other political aims. The politician hopes to energise his base with a bold-faced lie, while also changing the political subject matter to the fact that he lied, rather than other policy announcements made in the speech. The manipulative intent is still generic in the sense that the politician does not know who specifically he is influencing. However, it is less generic as he has grouped them into vague categories of "supporter" and "detractor".

Another challenge arises in examples of mass manipulation when a person appears to be subject to the influence and *is* influenced but was not the target of the manipulative intent. The politician may have grouped his audience into two vague categories, but there will be people who are neither. Furthermore, the politician will have theoretical knowledge that this is the case. What if these people are influenced and believe the lie? Did the politician have manipulative intent towards them? One way to understand this is to remember that the categories are very vague indeed, and the politician does not know who fits into what category, if at all. Perhaps we had better say that there are those he thinks will believe the lie, or could do so, and those he hopes to distract if they don't believe the lie.

Intent is complicated, and I do not think we need to have an extremely specific intent for each context like the above examples. Rather, it is enough to understand that manipulative intent can be generic or less generic, with specific restrictions. It is enough to say that the manipulation as trickery account, which requires the influencer intentionally influence their targets in a manipulative way, can do so with little knowledge in complex or less complex ways. This is sufficient to say that a manipulator does have manipulative intent, even if it is to an audience of millions they have never met and only assume to be rational agents capable of being receptive to the influence.

There is one other important question we cannot leave unanswered. What if I explicitly do not intend to manipulate one agent or group of agents in a mass audience, but nevertheless some of those are influenced in a manipulative way? To explore this, we can use another example.

[88] Conditioning is also a form of manipulation which, even if you know it is happening, can still influence you.

Alcohol-Free Trickery: Jim is a young teenager and is fed up with his schoolmates bragging about how much alcohol they can drink when he thinks they're overexaggerating. He organises a house party where he will offer a punch bowl to his guests. He loudly proclaims that the punch is very alcoholic, when actually it isn't. Jim wants to watch his guests make fools of themselves acting drunk, when actually they haven't drunk any booze at all. He is tricking them. Jim tells only one other person, his friend Bob, that the punch isn't alcoholic. However, he is astonished to see that Bob is acting just as drunk as the rest of his friends. It turns out that Bob actually forgot his conversation with Jim a couple of weeks ago concerning the ploy and was influenced by Jim's lie.

Jim intended to lie to everybody at the party, a large audience which includes Bob. However, Jim did not have manipulative intent towards Bob. Nevertheless, Bob appears to have been manipulated. Thus arises a challenge to the trickery account. The trickery account holds that manipulative action must be intentional by definition. Yet this example appears to show that Bob has been manipulated by Jim accidently, without Jim intending to do so. I'm happy to accept that Bob was manipulated without Jim's specific intent, but I hold that Jim still did have intent towards Bob. Jim just wasn't aware of it.

Jim intends to manipulate a set of people at the party, call this set A. Set A are those agents at the party who do not know that the punch has no alcohol in it. Jim thinks this does not include Bob, but it actually does. Therefore, as Jim intends to manipulate set A, he also manipulates Bob with intention. This is despite the fact that he did not consciously intend to manipulate Bob. Jim does accidently manipulate Bob, but this does not entail he did not subject Bob to influence without intent. Rather it is a consequence of a broader intent that Jim did not fully understand. Jim did not know the full consequences of his influence. We can contrast this with examples of pure unintentional influence. If you lie to somebody, intending only they be deceived, yet unbeknownst to you another agent was listening to your conversation and was also deceived, you had no intent whatsoever towards them. Jim does have intent towards Bob, as part of his intent towards a more vaguely defined group.

We could imagine if Jim were to sarcastically state something he believes is false to Bob, and Bob misses the intent for the statement to be sarcastic. Jim might say "Yes Bob, you should definitely go to the party wearing *that*" intending to be sarcastic, while Bob might take this as a genuine recommendation. Jim did not intend to deceive, quite the opposite. Therefore, we cannot characterize him as acting manipulatively, even though Bob has been made less ideal by Jim's words. The difference between the two is that Jim had intent towards the group, and by proxy Bob. A similar case could be if Jim was an arsonist and set fire to a building he thought was abandoned, while in reality it does have a person living in it. Jim did not have intent to burn alive that specific person, but did intentionally burn the building down. While I would not go as far as to say Jim the arsonist's intent is the same as if he definitely did know the person was in the building, I'm content to say that the person was killed intentionally. Therefore, I'm content to say the intent Jim has in **Alcohol-Free Trickery** is sufficient to say and the that we do not need to revise our definition further. It is a quirk of imperfect knowledge and generalised intent that such actions can have unforeseen consequences for the influencer. Of course, if the influencer wants to mitigate this, they must make their intent more specific by collecting more accurate information to base their intent on.

What this example highlights is that we would be remiss to think that manipulators are aware of the exact limitations of their intent. As in previous chapters, there are examples of people who only realise they acted with manipulative intent when they reflect on their behaviour after the fact. The preacher in **Apocalyptic Preacher** may believe they are truly acting sincerely at one stage of their career, and later as they acquire more and more money, transforming their preaching to extract

more money from their followers, reflect and discover that their true intention was manipulative from the start. Crucially, the manipulation as trickery example can survive the problem of the undefined audience if we understand generic forms of intention can suffice to meet the condition of intent required by the trickery account.

The Undefined Manipulator Problem

The trickery account relies on us knowing the intent of the agent performing the influential act. Without knowledge of this intent, we cannot conclude definitively whether the act is manipulative or not. Of course, I have already remarked that we don't have perfect knowledge of the intention of any agent, we rarely know our own intentions with perfect accuracy. So, we have to examine an agent in context. We watch the preacher's sermons, and how harshly they advertise their survival goods to their flock of people who have just been told the apocalypse will come any day, and make our own inferences as to their intent. Nevertheless, if we did have perfect knowledge, we would of course know the intent of the influencer for certain. In this theoretical world we have some interesting challenges to solve when it comes to influential acts. Specifically, media which influences others that is certainly the product of one or more agents.

We have not yet asked whether a piece of media can be considered 'manipulative'. It would certainly be natural to describe media, for example war propaganda produced by a government, as manipulative media. I do not think by calling a piece of media such as a poster manipulative we are saying the poster itself is acting as a sort of agent. Rather we are stating that we believe the agent who made the poster had manipulative intent. On this understanding, a piece of media can only be described as manipulative by reference to the intent of the person behind it.

One question is at what point the media becomes manipulative. More specifically, at which point does the manipulative act occur. We could first consider that it occurs at the moment of creation of the media. If I write a letter to a person intending to manipulate them, then we would say I act manipulatively when I write it. However, I could go through multiple drafts of the letter, and then change my mind about writing it altogether. Similarly, a manipulative statement, spoke aloud, is also 'media'. We would not say that I act manipulatively when I think of the statement, if I never actually speak it. I might form the intent, but the act of thinking manipulatively is not equivalent to acting manipulatively. We can instead consider the publication of the media, its expression or communication. This would be akin to speaking aloud, communicating in a manipulative way. It is important to note the communication need not have any actual effect on another agent to be a manipulative act. If I try to lie to somebody, but they do not hear me clearly, then I acted manipulatively even though I failed in *actually* manipulating them. Attempted manipulation is acting manipulatively, regardless of success.

Let's test this with an actual example.

Spiteful Review: Will used to be friends with Harry, who owns a restaurant. They've since become enemies of one another. Will decides to submit a scathing review of Harry's restaurant in their local newspaper, hoping this will harm Harry's business. However, Will is a terrible writer, so he decides to hire a professional writer to write the review for him. He instructs the writer to write a review which paints the restaurant in a bad light. Will then sends the paper to the newspaper, who publish the spiteful review.

Three agents are involved in this example. Firstly, we have Will himself, who organises, directs the content of, and pays for the article's creation and publication. Secondly, we have the writer, who creates the article and is paid to do so. Thirdly, we have whoever is in control of the newspaper, let's

say the editor of the newspaper, who consents to publish the article. Without question it is the intent of Will which we should be most interested in. Without his intent which directs the article's creation, how it is created, and how it is published, there would be no article. Yet he neither creates it directly, nor performs the act of publishing it. Will is one-step removed from the process, just as a mafia boss would be one-step removed from a murder he directed. He caused the murder, but did not perform the act of murder directly. Of course, we can still say that Will acted manipulatively, he just used other agents as tools to do it. If I shoot a person in the head, it's the bullet which did the killing, but I fired the bullet, so I did the killing in the sense that matters. As written, we can assume that the writer and editor are hapless tools used by Will, the manipulator. Will could tell the writer he really did have an awful experience but just cannot put it into words. The editor could have no knowledge of the restaurant and see the review as a sincere reflection of Will's awful experience. We could argue that the editor has a duty of care to not publish falsehoods, however we can equally assume for the sake of the example that he did what due diligence he could be reasonably expected to do. Consequently, Will is the manipulator because it is his manipulative intent the media embodies, not the intent of the writer or the editor.

However, we can also alter the example to arrive at a different conclusion. Perhaps Will tells the writer in his commission that he wants the writer to produce an article which lies about Harry's restaurant. The writer is no longer just a tool, but an agent who must choose whether to intentionally write the article which they know will be used to influence others to believe falsehoods. Thus, if they write the article, they will be writing with manipulative intent. They don't publish the article, but they do know it will be published, or at least that Will intends to do so. Compare this to a friend asking you how to lie to their spouse about their affair, because they don't know how to phrase it. If you give them the words, knowing they will tell their spouse them, then you are acting with manipulative intent in a way that is different from only constructing a manipulative statement in your head without actually speaking it.

It appears then that the manipulative act is not just found in the act of communicating/publishing the manipulative media, but in creating it knowing or suspecting it will be published.

The editor likewise may be told that the article contains falsehoods, exaggerations or will cause inappropriate emotions in the reader. If they know this and publish it, they will be acting with manipulative intent. Compare this to a person being told to lie to another person, being given a script they should read. If they know it is a lie, but repeat it, then they act with manipulative intent. For both the writer and the editor, we could also say that they know the media contains falsehoods due to their own knowledge. Perhaps they've both been to Harry's restaurant and know that what Will wants to be written is misleading or false. The same standards apply.

We're in a strange philosophical puzzle here. We must accept that an agent can act manipulatively despite neither creating directly—or publishing directly—a piece of media which will influence others manipulatively. If the creator and publisher have knowledge that it is manipulative, they also act manipulatively despite only creating, or only publishing the media. This does not apply to a person manipulating alone. If Will wrote the article, but decided not the publish it, intuitively he did not act manipulatively, just as a person doesn't lie by constructing the lie in their head. At least with the publisher our intuitions are more satisfied. A publisher might happily publish genuine opinion they disagree with if they think that the intentions behind the piece are sincere. If they know it is insincere, then they do act manipulatively by publishing it, and I am satisfied by that conclusion.

To resolve the dilemma, we should consider that Will are the writer do not just act alone, but act collectively, performing different elements of the act of manipulation itself.

Acting Collectively

There are a variety of theories about groups acting collectively or acting as a single supervening group agent. There are many accounts of different sorts of groups who may share different sorts of aims, and what conditions are necessary to say they act together, not merely coincidentally at the same time. There are even those who believe that there is no such thing as group agency, only individual agency. It can be natural to ascribe group agency to various teams who work collaboratively on a single work. If friend and I build a piece of flatpack furniture for our new flat, we can look at the finished product and say that *we* assembled it. It falls to the philosophers to work out just what we mean by the statement. Crucially, I should note that this project is not concerned with whether theories of joint or group agency are correct. We are only concerned with whether or not these theories, if true, can aid in explaining how the trickery account can apply to examples of mass-manipulation. If there is no such thing as coherent group intention and action, then we must fall back to analysing such actions relative to only the individual. With that limitation, let's examine the potential of group agency to assist us.

Firstly then, we should remind ourselves of what it means for an individual to act manipulatively. For an individual to act manipulatively, they must a) intentionally influence a target agent *i*; b) intend that the influence against *i* cause *i* to fall short of one or more cognitive ideals the influencer believes govern *i*'s beliefs, desires or emotions; or c) lack an intent to avoid causing *i* to fall short of those ideals.

From this we can understand what abilities a group must have to act manipulatively. For a group to act manipulatively a group must be capable of both intending as a group and holding beliefs as a group. It must not only influence purposefully, but with the intent to make the target fall short of ideals, or lacking an intent to avoid it. Consequently, it must be able to hold beliefs about the cognitive ideals of the target agent, in order to make the required judgment about the influence it is about to enact.

The actors we have are Will, the writer, and the editor. For this sort of small group, we can use Michael Bratman's theory of shared intention. In his article he develops a definition of what it means for two people to intend to *J*, where *J* is some action.

"We intend to *J* if and only if

1. (*a*) I intend that we *J* and (*b*) you intend that we *J*
2. I intend that we *J* in accordance with and because of 1*a,* 1*b*, and meshing subplans of 1*a* and 1*b*; you intend that we *J* in accordance with and because of 1*a*, 1*b*, and meshing subplans of 1*a* and 1*b*.
3. 1 and 2 are common knowledge between us." [89]

By 'meshing subplans' he means plans individuals have to do the act *J* which do not conflict with one another. For example, some friends and I might plan to attend a concert together. This does not mean we both need to intend to attend the concert with the same steps such an intention may entail. Those steps may be travelling to the location the concert is being held at, buying tickets etc. Rather, we will have plans for ourselves as individuals as to how we are to attend the concert which must not conflict with each other. If I intend to drive, while my friend intends to take the train, that does not necessarily produce a conflict which would entail *we* don't go to the concert as a joint

[89] (Bratman, 1993, p. 106)

action. We must both intent that we act with the intention we act together, and that we intend to do so with meshing subplans.

An example is given by another writer, Debora Tollefsen, demonstrating the theory. She asks us to consider the common scenario of two people who intend to jointly wash the dishes after a meal.

"Participants must intend that their subplans mesh with each other. This doesn't mean they have to match. If my subplan involves the intention to wash the dishes with a certain kind of soap, you intend to wash them with hot water, and I have no preference about the water temperature, then our subplans mesh even though they don't match exactly. But if we have subplans to wash the dishes with different water temperatures, our subplans do not mesh, as it would be impossible to accommodate both."[90]

I would state that the conditions above do obtain in the case of Will and the writer. By commissioning the writer, Will asks the writer to act jointly with him in the creation of the media. The writer will write the words, Will provides the topic and its intent. Moreover, they both have meshing subplans concerning the division of labour required to bring the article into being. However, if Will does not make clear that he has manipulative intent, then he is not asking her to jointly act manipulatively. The writer would not have the knowledge required to satisfy any of the conditions. She would not intend to manipulate, meaning 1b would not obtain. Thus, 2 and 3 would fail automatically. The writer would not know she is writing lies on Will's behalf. They act jointly in the production of the article, but not in acting manipulatively.

What if Will does make it clear he wants her to write falsehoods? In that case, all conditions do obtain. She therefore acts manipulatively as part of a joint action with Will, they just take different roles in doing so. The same can be said for the publisher. When they publish the article, if they do not know Will's intentions, then they are acting jointly with Will in the act of influencing but not the act of manipulating. If the publisher does know, then they also jointly act manipulatively.

Neither the writer nor the publisher's actions would count as manipulative action in isolation. Will is the lynchpin who provides the article's intent. The publisher and writer only act manipulatively jointly, not singularly.

Of course, one could deny Bratman's concept of joint action. I believe the analysis mostly survives the denial. We could consider a situation where none of the actors actually communicate transparently with one another. Perhaps the writer works for a ghost-writing service, never actually interacting with Will. She simply writes what her employer tells her to write, with no knowledge of what it will be used for. In these circumstances she lacks manipulative intent. For all she knows, the article is for a work of fiction or a lazy student's assignment. If the brief came with knowledge of what Will intends to do with it, that he has manipulative intent, then the writer would be complicit. Even if we say that she acts only on her own, she acts on the intent Will provides. I note here that it does not matter when or not the article is ever published, only that the writer reasonably believes it will. This is sufficient for her to intentionally write to influence others in a way which is manipulative. The publisher similarly either knows of the intent of the writer or does not. They either share and spread what they know to be lies, or what they mistakenly believe to be genuine. Knowledge of the intent remains the most important feature, regardless of whether or not we want to accept joint action and state they manipulated jointly, or acted manipulatively in parallel to one another, not jointly in the special sense.

[90] (Tollefsen, 2015, pp. 34-35)

One objection might be that the writer or publisher may not share Will's intent. They do not harbour anger towards the owner of the restaurant which motivates Will. There may also be other material factors in play. The writer or publisher might not want to aid Will, but need to do so for money or other reasons. Nevertheless, they act to assist him intentionally, just not fully willingly. This may change our moral evaluation of them, but not the fact that they acted with manipulative intent.

The unidentified manipulator problem asks us who we look to for the intent behind an act of influence when multiple agents, perhaps with differing intentions, take part in the act of influence. With either the aid of accounts of joint action, or with a separate analysis which denies accounts of joint action, we can see that at least in regard to small groups, the intent which decides whether an act of influence is manipulative is the intent of the person who motivates the act. Those who aid in the act can be acting jointly, if we accept such a definition, and share the intent. Otherwise, those individuals can be ignorant and blameless, or collaborators, reluctant or otherwise, who mirror the intention of the motivator. This resolves the unidentified manipulator problem.

However, there are other accounts of groups, and group action, which examine larger, more established groups than those I have focused on before now. No account of mass manipulation would be complete without being able to explain what we mean when we accuse a government or a corporation of acting manipulatively. Indeed, on the grand scheme of things we might be more concerned with the conduct of these sorts of entities simply due to the power they wield. In typical speech, we don't just accuse any one individual who is part of these groups as acting manipulatively, but the group itself. A government at war may produce propaganda posters and be accused by its detractors of manipulating its citizens. Similarly, a company may be accused of deceptive marketing practises and held liable for it if the laws allow.

An immediate problem is that what we can call 'corporate groups' (which I will use to refer to all of these groups) are not thinking beings capable of the phenomenological experience of "intending to manipulate". Therefore, if we do want to say that a corporate group has manipulative intent, we will not be ascribing it the same sort of intent we would a person. However, it is also true that we can describe a company as having an interest in influencing people to be less ideal. A tobacco company has an interest in its customers believing that tobacco usage does not increase the chance of those users' developing cancer and other conditions which might not necessarily be shared with the director of the company or its board of directors.

I believe the analysis of smaller groups, without accepting the existence of joint action, can help us here. A corporation may be a much larger group of people acting either jointly or parallel to one another, but there will be some who provide the intent behind an act of influence and others which provide the means to carry out the act. If a board of directors or department committee is the brain of the corporation who decide to create an advertisement which they believe will influence viewers in an unideal manner, the marketing team who create it are the hands. Just like in **Spiteful Review**, one element of the group provides the intent, and the others produce.

We could go into further detail here about how decisions are taken in a corporate group. Tollefson references Peter French who talks of "CID", or the "corporate internal decision structure".[91] Phillip Pettit in '*Groups with Minds of Their Own*' provides a full account of how such groups can make decisions using various methods. Moreover, he explains the paradoxes which need to be resolved to make sense of such group action.[92] I do not believe it is necessary to explain those accounts here.

[91] (Tollefsen, 2015, p. 45)
[92] (Pettit, 2003)

Ultimately, we do not need to know exactly how a corporation can come to possess manipulative intent. Rather we can assume that somewhere in the minds of those that made the decision, they had such an intent. We already know that corporations are collections of people. Those individuals can have individual manipulative intent. We have already examined how groups of people can work together, whether we consider their actions to be joint or not, to act to influence people in a manipulative way. Therefore, how such deliberations take place or how the intent is formed is not our concern. When we accuse a corporate group of acting manipulatively, we are accusing the decision makers of that group, by whatever means we want to define them and how they make decisions, of holding manipulative intent which motivated their acts of influence. How we theorise the particulars will depend on what account, if any, of corporate intent we want to adopt. That is not the purview of this project.

On a practical level, it is key to bear in mind that we will rarely *know* the intent of individual influencers such that we can call them manipulators with absolute certainty. With corporate groups and other groups, this becomes even more challenging due to the number of people involved in creating the intent. To practically treat a company and its influence is manipulative then, we must consider those interests we know corporations can have. With this approach we can explain why we often evaluate the influence of corporations and governments as manipulative, despite the fact that the picture of what the decision-makers and thinkers of a company actually believe about their influence is so remote and unknowable to us as outsiders. If a company is a for profit entity, then its first and we might say 'primary' interest is the acquisition of profit. Though anecdotal my perception is that there has been a resurgence of "cause marketing" advertisement strategies by corporate groups as of late. It is especially true that such marketing campaigns have 'gone viral' and dominated online discussion of the brand and whether or not such marketing has a positive effect on society. One example would be the Gillette razor brand's '*We believe: the best men can be'* advertisement which sought to connect the brand's traditional promotion of masculinity to the sentiments of '#MeToo' movement.[93] Though most of the controversy was certainly motivated by anger and disgust at the socially progressive message of the advertisement itself, there were also accusations that the advertisement was a cynical attempt to capitalise on a social movement to sell razors.

Ultimately, Gillette, and other brands, do not exist to promote social change. They exist to sell a product. Companies may see themselves as having a broader social responsibility in addition to their primary function of being a profitable business; after all, corporate groups are made up of agents with their own moral responsibilities. They may not pursue social responsibility if it didn't *also* produce profit, or at least not have a detrimental effect on the company, but this does not mean they are necessarily insincere in attempts at positive social change. Additionally, whether their intent is insincere, and thus manipulative, or not, it does not necessarily detract from the actual positive social change such influence may have. However, accusations of insincerity can be justified on a case-by-case basis. The subjects of such influence will need to use their own knowledge of the incentives a corporate group and its decision-making members may possess, as well as evaluate their behaviour in other avenues of business. For example, a company may advertise in the cause-marketing style to claim they are in favour of promoting equality between the sexes, yet at the same time have almost all male corporate decision makers. This will aid the inductive process of evaluating whether such influence was likely made with manipulative intent, or if it was sincere.

Certainly, cause-marketing likely will not be going away any time soon. One meta-analysis of industry and research literature on cause marketing found that customers self-report being "more

[93] (Gillette, 2019)

likely to switch to brands that support a cause".[94] However the same meta-analysis notes that "23% of consumers said that the primary motive of companies is a tax write-off and another 20% said it was publicity".[95] Consumers may appreciate if their purchase is in some way contributing a charitable cause, whilst still harbouring suspicion that the company is sincere.

Artwork, Media and Manipulation

At first glance, we might not be able to point out too many differences between an advertisement styling itself as a 'short film' and a short film shown at your local art cinema. Of course, there are differences. One is made to sell a product, the other to prompt debate about one topic or another. One may directly address the audience at some point in order to explicitly convey a message, whereas the other might simply present its subject matter for the audience to react to. Yet both use actors, sets, costumes etc. Both are examples of primarily visual media. Both seek to influence the audience. Therefore, at least in principle, the work we have already done regarding the undefined audience problem and undefined manipulator problem can apply to visual artworks such as film and television programmes. Furthermore, it can apply to other artworks. A novel can be manipulative, in the same way an opinion piece can be in your local newspaper, and so on and so forth. We can understand them just the same and need not trouble ourselves with questions of joint creation of manipulative artworks and so on.

However, that is not to say that we have nothing further to be concerned about when it comes to artwork, media and manipulation. On the contrary there is a general worry when it comes to accounts of manipulation that they may be overinclusive when it comes to art. This is especially true for accounts such as the trickery account, which I argue defines manipulative acts as immoral by default. I believe this worry is motivated by a central intuition that film, novels etc evoke emotions seemingly without our active participation or justification. The second element of this intuition is that the fact that artworks do this is not ethically problematic. Whereas an overinclusive and moralised account of manipulation may say they should be.

Consider the use of soundtracks, audio accompaniments to films which almost always contain music. When the director wants the audience to feel sadness, a suitably sad song plays. When they want us to feel hopeful, the music changes to match. One might claim that because the film is attempting to influence our emotions, it is manipulating us, eliciting those emotions from us involuntarily. Additionally, we might argue that feeling emotions are not justified when concerning fictional scenarios. When a film makes us feel fear, one could argue we have no rational reason to feel afraid. We are not actually in danger. Thus, the film knowingly provokes an emotion which is unjustified.

The intuition that eliciting emotions in an audience, in general, is manipulative, is one I reject. While the use of the term is natural due to the analogies with direct manipulation of say, a marionette, or an instrument, metaphors discussed previously, we should move away from this metaphorical use. The more precise statement is that media can influence our emotions. There is nothing inherently objectionable about our emotions being influenced, even involuntarily. If we see a person in tears, we should not claim that they have manipulated us, simply by presenting us with an image we may react to emotionally. We can want to avoid feeling an emotion, but feel it anyway, without this being evidence of manipulation on the part of whatever phenomena elicited that emotion from us. We may even feel emotions which in that context, result in us falling short of one of the ideals for emotion. Yet this does not mean that it was the intention of the agent eliciting the emotion. Even if they are intending to influence us, a stressed and upset customer trying their best to get us to break

[94] (Rozencher, 2013, p. 183)
[95] (Rozencher, 2013, p. 184)

company policy for example, this is not sufficient to declare the influencer manipulative. They may think that we, as the target of their influence *should* feel sympathy, or sadness towards them. It depends on whether they have manipulative intent, just like another other instance of influence.

The second worry when it comes to fiction is that if a film or other media is fictional, it cannot invoke emotions with justification. We can be annoyed that a book has written a character a particular way, that can be justified. However, we cannot be justified in feeling sad that a character has died *because the character died*, as the character is not real. A character cannot die as they never lived in the first place. This problem is the paradox of fiction. The premises of the paradox can be stated as:

"(1) that in order for us to be moved (to tears, to anger, to horror) by what we come to learn about various people and situations, we must believe that the people and situations in question really exist or existed;

(2) that such "existence beliefs" are lacking when we knowingly engage with fictional texts; and

(3) that fictional characters and situations do in fact seem capable of moving us at times. "[96]

We are not looking to question the fact that films really do evoke what appear to be emotional responses. Rather the argument regarding manipulation looks to premise (2), we do not have rational justification for the emotions which are invoked, and thus purposefully and invoking such emotions in an audience counts as manipulation.

We do not have the space to delve into the philosophy of fiction, as interesting as it is. Suffice to say that the many attempts to resolve this paradox may well be compatible with the trickery account, should they survive the debate on their own merits. However, I wish to step around this argument by turning back to our previous discussion regarding the **Surprise Party** example, in Chapter 3. The result of our discussions was that both explicitly and implicitly, a person can consent to influence we would ordinarily believe is manipulative, so long as it is limited in scope and necessary for attaining one of our aims other than maintaining our beliefs, desires and emotions in an ideal way. When voluntarily viewing media, engaging with it etc, the viewer may consent to be manipulated, meaning it is no longer manipulation. Rather than thwarting our attempts to meet the ideals, it aids us in the broader sense of achieving them. Of course, this is not a line of argument we need to use if an appropriate resolution to the paradox of fiction is adopted.

We can borrow from the philosophy of fiction to solve some more periphery concerns to do with the interplay between the trickery account and fictional media. Some stories use unreliable narrators, or otherwise lie to the viewer within the context of the fiction. This can be solved either be appropriate adoption of a philosophy of fiction,[97] or by understanding the viewer has consented to it.

To summarise then, I believe we can place to one side worries that fictional media itself is mischaracterised in a problematic way by the trickery account. We do not need to worry that we need to take a moral position against techniques used in film making when it comes to artwork addressing fiction. Either the philosophy of fiction, with its many comprehensive accounts, can properly explain how we can separate fictional truths, and emotions reacting to fictions from real

[96] (Schneider)

[97] My personal preferred philosophy of fiction is Kendall Walton's. Best expressed in his book *Mimesis as Make-Believe* (Walton, 1990). Under his account of fiction, lies told by an author would be fictional lies, part of a game of make-believe. Fictional lies are not the same as real lies, and do not necessarily share the same moral analysis.

truths and emotions elicited by real phenomena, or we can consider this influence consented to, and thus not manipulation proper.

Fiction and Reality

Of course, fictional media does not restrict its influence to only fictional topics. More often than not, fictional media aims to influence us regarding reality. For example, many pieces of historical and contemporary media contain moral messages, are metaphors for real events, or contain other commentary on the real world. The artwork does not only present us with fictional scenarios, but want us to compare these to the real world, to change or reinforce our beliefs, desires or emotions. An example would be the extremely controversial film '*The Birth of a Nation*', adapted from the novel '*The Clansman*'. The film portrays black people as physically and sexually violent; and presents a narrative of the civil-war and reconstruction era which is at odds with the objective historical record. We must be careful not to instinctively take an objective approach to manipulation when considering such films. A film can be condemned for miseducating its viewers, or for being otherwise morally repugnant, while not being manipulative. The directors can be sincere in their objectively false beliefs. This does not mean they cannot be manipulative in the manner in which their beliefs are argued for, as previously discussed in the **Sincere Cult Leader** example in Chapter 2.

I would argue that the context of watching a film, or otherwise engaging with a piece of influential media which is primarily fictional, does not include a consent to be manipulated in regard to feature of the real world. If there is a moral message the film wishes us to consider in its story, we would expect it to be sincerely put without the use of manipulative techniques. Similarly, if the media intends to educate us about the real world via the medium of fiction, we would expect them to be sincerely expressing what they think to be truths, not falsehoods. Though there is some room here for the limitations of what can and cannot be expressed, or simplifications of some facts. I am thinking particularly here of programs which are aimed at children which use fictional content as an educational medium. Artists appear entitled to expect some engagement from their audience on such issues, for an audience to understand what is trying to be conveyed. With media literacy comes an understanding from the audience as to what is entirely fictional influence and what is influence the audience is asked to apply to the real world. The exact limitations of this, I do not know.

The fact that artworks are set in historical time periods can create some interesting conundrums for the account along these lines. Remember, it is not enough for an agent to not intend to make their target less ideal, but they must not manipulate through bullshit either. They must have a regard for the truth and other ideals of those they influence. Consider then the variety of films and novels which are set in a historical time period but do not portray these periods at all accurately. In the online sphere of medieval history enthusiasts, Hollywood tropes such as firing flaming arrows in battle, or firing arrows in synchronised volleys, are routinely mocked as not being historically realistic. Yet, for many viewers of these films who do not know anything of note about medieval history, myself included, may take these inaccuracies as a genuine feature of the period.

We may consider that a consequence of the trickery account is that creators of fictional artworks which place themselves in a historical setting do need to at least pay mind to whether or not they are misleading their audience as to features of the setting. At the same time, it strikes me as neither reasonable nor desirable to burden artists with also being historians. Again, artists have a reasonable expectation that their audience will be able to tell the difference between what in a setting is meant to be a realistic portrayal of real events and what is artistic liberty. Therefore, I conclude that we should not be overly concerned with this. The intent of the artists in this case will almost certainly be that audiences should not take from their art facts about the setting which are falsehoods. The artists are not being negligent in regard to their influence of the audiences' beliefs, desires or

emotions. To accuse them otherwise would be like accusing a person being sarcastic of misleading the person they address, if the person they address is particularly gullible or otherwise takes the sarcastic comment literally. The audience is undefined, and we have already stated that influencers address the vague audience with some basic expectations. In the case of artwork, that includes a basic media literacy. Of course, I imagine at the extremes, depending on the artwork and how its portrayed, artists could stray into acting manipulatively.

Reality Television, Fiction and Non-fiction

If we can say as much about fiction, we may conclude we can be much stricter when it comes to non-fictional media. With only one context to consider this should be straightforward. If a documentary intentionally influences us in a way which makes us less ideal, then it is manipulates us the same as a news article or historical manuscript.

The difficulties become more apparent when we blend the two mediums. 'Reality television' as a genre, straddles the line between fiction and non-fiction media. An example would be the television show '*Love Island'*, which as of writing is still running. In this program, contestants compete for a grand prize. They are isolated with their fellow contestants on a set, in this case a Mallorcan villa. They must perform various tasks relating to their choice in romantic partner for the week. Contestants must 'couple up' on the first episode and can then 're-couple' throughout their time on the show. Love Island's episodes are mostly complications of the clips taken of the islanders living on the island and interacting with each other. The majority of the entertainment on the show is gained from the spats, drama and arguments between the different couples. Each week, contestants are 'voted out' by the public, or do not find a fellow contestant to 're-couple' with, meaning they are eliminated.

Programs like this are interesting to us, because they present a mix between fiction and non-fiction. On one hand, we are presented with real people and real events, not fictional ones. The contestants have real relationships with each other, commit real betrayals of one another and have real arguments. Otherwise, it wouldn't be as entertaining. Yet at the same time these situations are manufactured, then presented to us in a manufactured way by editing. The contestants themselves are carefully selected by the showrunners to be entertaining people to watch. They are also forbidden from talking about anything other than what is happening on the island. One contestant explained in a BBC interview:

"So if me and you say: 'Let's talk for three hours on the evolutionary underpinnings of why men fight each other,' they'll come over the tannoy and say: 'Islanders, we can't talk about the outside world, please.' What they want is you to talk about Omar [a fellow contestant], and the girl who mugged him off last night."[98]

The show is edited down to an hour from what could be days of footage. Additionally, 'confessionals' are filmed, where contestants speak directly to a camera in a private room about their experiences. From reality, the producers of the show are able to construct narratives. It is in this narrativization that a fiction is presented to the audience. The showrunners take messy, boring reality and turn it into something which can be understood and absorbed by an audience in as little as a few minutes or less, considering the number of contestants and companion narratives which need to be presented.

The issue is that by taking non-fictional content and organising it so that the showrunners can present a fiction, they create a narrative which could be misleading, misrepresenting what occurred

[98] (British Broadcasting Corporation, 2022)

and what sort of people the contestants are. At the extreme, real people are shown as caricatures. They are reduced to tropes, easily presented, and easily understood by the audience. Manipulation can occur by influencing audiences to believe falsehoods, or to feel inappropriate emotions, as well as by making salient elements of the television show, and thus its contestants, which should not be.

We could try to say that we must have faith that the audience understands this. The audience should understand they are being shown an edited, manufactured narrative from real world events. Yet I am sceptical that this defence can be applied here, as it could in examples of pure fiction. Reality television is so compelling because it has an attachment to the real. It is so interesting *because* the people shown are real, and the events are real. People form judgments of the people as they are shown on the programs.

Perhaps then we have another blurry line to draw, as unsatisfying as it is. The fiction could be accurate, more or less, to what occurred and how contestants or other subjects of filming acted. There is again a degree of artistic license the audience must be taken to understand by the program's creators, absolving them of accusations of manipulation in that respect. Yet this format is ripe for accurate accusations that the fiction created departs significantly from reality, while purporting to be accurate with minor alterations.

In summary then, reality television blends fiction and non-fiction in a way which can easily lead to accurate occurrences of manipulation by the creators of such shows. Due to the demands of the format and a split interest between accurately reporting what occurred and making what occurred appear more entertaining and understandable, showrunners run the risk of intentionally misrepresenting real events, thus manipulating their audience in one manner or another.

Conclusion

Mass-manipulation is a challenge for the trickery account, but it is a challenge which can be overcome through careful examination of the problems and pointed application of the account and its broader themes. The trickery account cannot apply to an example without knowing where the influence originates, and the intent behind it, as well as who the audience are, such that the intent can be properly formed. Examples of mass-manipulation challenge the account by giving examples of manipulation performed to theoretically all of humanity at once, of which the manipulator could not possibly have knowledge of all of them, as well as examples where the manipulator's themselves are many, all with their own thoughts, desires and emotions.

The unidentified audience problem can be resolved by understanding that manipulators can have intent, which is generic in form, aimed at people based on singular traits. We accept that agents can have such intents, and I have shown that these intents are suited to the trickery account. Whilst we can complicate the picture by introducing split intentions and the actual audience the manipulator addresses, we can still make sense of intentions directed at multiple, vague persons which can be manipulative.

The unidentified manipulator problem can be solved by appeal to theories of joint action, but I have argued we don't necessarily need to do to. Intention can be provided by a single agent then mirrored or shared by others who perform the actual work of producing and publishing a manipulative act of influence. On both small scale and large scale acts, we can identify, at least theoretically, the agent or group of agents which form the manipulative intent, however it is then enacted.

Finally, we have looked at the problems of artwork and media. I believe we can reject the idea that fiction itself is manipulative by appealing to existing accounts of fiction, the truth value of

statements regarding fictional subject matter, etc. However, it remains that non-fiction and the blending of fictional and non-fictional influences can produce some circumstances where the techniques of media, such as editing choices, can be done with manipulative intent and thus constitute manipulative acts by the publishers of the media.

Certainly, more could be written about these topics. I do not pretend to have solved every conundrum. What is crucial is that I believe I have shown the trickery account *can* be applied coherently to these sorts of examples, vindicating itself as not an account for a limited set of circumstances alone, but aiding us in identifying and understanding manipulative acts in broader contexts. This is especially crucial when we consider that it is these acts of influence which can have the most power in a world where practically speaking, anybody with a smartphone and an internet connection can influence millions with the touch of a button.

Chapter Five: Manipulation and Design

Introduction:

In the previous chapters, we've thoroughly examined the ways agents can directly manipulate each other. Unique challenges were provided by examples where an agent manipulates a large audience in a more remote fashion, or when we identified a collection of agents, rather than an individual, as acting manipulatively. The final broad category of interactions I've identified which produces unique challenges for any account of manipulation is manipulation through design. This is a category which, like mass manipulation, an account of manipulation must address for it to be complete.

The design of systems, interfaces and spaces can influence our behaviour just as significantly as a shouted command or a carefully put argument. In this chapter we will consider how intentional and unintentional design choices can potentially influence us, as whether that influence can be manipulative. This is as much as exercise in the application of the trickery account as it is an exercise in the further development the account. Some elements of this chapter will serve to show how the trickery account, as it has already been developed, can be applied as an ethical and practical guide. Other elements will challenge the account, so we must resolve these challenges in one way or another.

The first and most significant section will look at 'nudge theory', which broadly looks at how design principles can influence the designed space or system. Secondly, I will examine the use of social media algorithms and the design of interfaces. Finally, I will examine manipulation and social media.

Nudge Theory: What is a Nudge?

Behavioural economists Richard H. Thaler and Cass R. Sunstein begin their book '*Nudge: Improving decisions about health, wealth and happiness',* with a compelling and now much discussed example of school cafeterias.

Carolyn, who runs the catering program for a large city school system, conspires with a friend who works with supermarkets to alter the design of the various cafeterias to experiment with how they can influence the eating habits of its students. They find that, by changing which foods are at the children's eye height, which are presented first and which last, and in what location in the cafeteria they are presented, they can increase consumption of particular items by as much as 25%. Carolyn, Thaler and Sunstein explain, is a choice architect. Without necessarily changing what choices the children have, but how they are presented, they can influence the children who use the cafeteria. The choice architecture is the context in which we can make a choice, akin to how physical architecture like buildings, parks etc are the context in which we occupy and act within physical space. Like real architecture, the choice architecture is the framework in which we find a choice presented to us. Of course, there does not need to be an actual architect for there to be choice architecture. Some choices occur naturally without intervention from an agent. We can describe a hillside or a cliff as natural architecture. Just so, there is natural choice architecture which arises organically without an agent's involvement. Additionally, we can be our own choice architects, changing our current or future choice architecture ourselves. [99]

Thaler and Sunstein define a nudge as a way of designing choice architecture:

[99] (Thaler & Sunstein, 2009, pp. 1-2) this whole paragraph is paraphrased.

"A nudge, as we will use the term, is any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economics incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting the fruit at eye level counts as a nudge. Banning junk food does not"[100]

A nudge is therefore a form of intentional influence. It is an intervention; purposefully performed with the intent to affect which choice a person makes. It alters the context in which their choice is presented to them. There are a few more notable statements Thaler and Sunstein make which we should pay particular attention to.

The first is that choice architecture is not neutral. [101] There is no way to formulate the layout of the cafeteria so as not to affect the choice of those who use it. Rather, there are different approaches a person can take in order to affect the choice in different ways. Just as we are inevitably influenced by the design of buildings we use; we are inevitably influenced by the choice architecture surrounding a choice. In relation to the cafeteria example, Thaler and Sunstein argue that Carolyn *has* to influence the choices of the students in some way. Even if she directs that the cafeteria staff arrange food items randomly, that would count as form of intentional influence, nudging the students in whatever way that randomness happens to entail, for better or worse.

The second matter is how Sunstein and Thaler limit their definition of a nudge so that it does not include cases where there is a significant alteration of the 'economic incentives' of the chooser. In another paper, '*The Ethics of* '*Nudging*'', Sunstein emphasises this. He states: "A subsidy is not a nudge; a tax is not a nudge; a fine or a jail sentence is not a nudge. To count as such, a nudge must fully preserve freedom of choice. If an intervention imposes significant material costs on choosers, it might of course be justified, but it is not a nudge."[102]

There is some complexity to this topic, which Sunstein acknowledges.[103] It will be hard to draw a hard line between what counts as a 'significant material cost' and what does not. One example could be a government mandating a minor cost of say, ten pence to the price of disposable plastic bags provided in shops. The aim of the government is to nudge customers into reusing the plastic bags to reduce waste. Yet it is unclear whether this counts as an economic inventive like a tax which would not be a nudge. After all, to some this will be an extremely inconsequential cost, while to others in dire economic circumstances it could be quite significant. There may be no hard, categorical boundary to use to distinguish between changes to choice architecture which are a) 'soft paternalism' (imposing insignificant material costs) and thus nudges, or b) 'hard paternalism' (imposing significant material costs) which Sunstein does not want to define as a nudge.

For our purposes, it does not particularly matter whether we adopt Sunstein's limitations on the definition of a nudge. We do not have to draw these lines. Rather, it is enough to say that a nudge is an intentional form of influence. Therefore, like other intentional influence, the intention could in some examples be manipulative. Ergo, the act itself would be an act of manipulation.

Regarding the discussion of 'soft' and 'hard' paternalism I am reminded of cases previously discussed regarding the difference between coercion, and accounts of manipulation as 'pressure' discussed in the first chapter. We can definitely influence people's choice architecture by coercion. In the very

[100] (Thaler & Sunstein, 2009, p. 6)
[101] (Thaler & Sunstein, 2009, p. 3)
[102] (Sunstein C. , 2015, p. 417)
[103] (Sunstein C. , 2014, p. 57)

first chapter we considered how a highway robber who holds you at gun point doesn't, on one view of liberty, rob you of the choice to not give him any money, the robber just makes it an unacceptable option by introducing severe consequences for taking that choice. The robber alters your choice architecture in an extremely significant way, but technically the choice remains available to refuse the robber and take your chances dodging bullets. This shows that changes to choice architecture can and are subject to the categorisations of other methods of influence, and the concepts and definitions of those fields. We are concerned with applying the account of manipulation as trickery to nudges, so it is enough to say a nudge is a form of influence for that purpose. We do not care where the hard line is drawn, only that nudges are influence. What holds true for influence generally should hold true for nudges.

With these two elements in mind, we can consider whether *all* nudges are subject to the trickery account. Remember that the trickery account is a type of account of manipulation which accounts for occurrences where an agent subverts or bypasses the reasoning of their target agent. As previously stated, we can conceive of this as operating by eliciting a non-ideal response. We can also influence people by changing what the ideal response is for that person in that situation, such that the rational choice they should make is in line with our own desires. This typically requires we alter the choice architecture around an agent, rather than the agent's choice-making faculties themselves. These sorts of changes to choice architecture, if we do decide they are nudges, may not be the sort of influence the trickery account seeks to explain. If we did want to still identify them as manipulative, we would be using the term in the manner accounts such as the manipulation as pressure account use it, which is not the same concept. As stated in the first chapter, some cases can also be examples of both manipulation by ideal, and non-ideal response.

To summarise then, a subsidy on the price of a train ticket, aimed at increasing the use of public transport, if it is a nudge, might not be the sort of influence analysable under the trickery account. However, other accounts evaluating methods of influence may be able to analyse them in their own terms.

Consequently, I believe we can take the view that as not all nudges can be manipulative, accusations that all nudges are problematic because they are all manipulative must be flawed. It is this accusation, among other more nuanced concerns, which we will now move onto.

Salience Nudges and the Trickery Account

The accusation that all nudges are manipulative can be called the 'manipulation objection'. The objection comes in a variety of forms and is motivated by the belief that nudges are a threat to our autonomy, are performed without our consent, or are performed covertly. These often relate to accounts of manipulation about which I've already expressed my concerns. However, the concern regarding autonomy is the closest to matching the concern with manipulation as trickery. So, it is applicable in principle.

As I've already stated, we cannot state that all nudges are manipulative following the definition provided by the trickery account, simply because not all nudges can be analysed with the trickery account. Moreover, I think the objection is all the stronger if we limit it to some nudges, or some types of nudges with a shared method of operation. The alternative is that which Sunstein adopts when addressing the issues of manipulation in relation to nudges. He states that while "No one should deny that at least some nudges can be considered as manipulative within ordinary understandings of the that term. I have emphasized that any action by government, including

nudging, must meet a burden of justification."[104] Essentially, Sunstein appears to believe that if some nudges are manipulative, they may be justified anyway. This mirrors my own analysis of the ethics of manipulation as trickery. If a nudge is manipulative, it means that it is wrong, but may still be justified by the context. It simply stands in need of that justification.

Overall, it appears that the best way to consider the ethics of nudges is to consider the ethics of paternalism first, rather than manipulation. We may be of the belief that a paternalistic nudge by a government office is not appropriate, regardless of whether or not that nudge is manipulative. This is of course outside the scope of this project. It is worth noting that I believe there are clear examples of nudges where our intuitions should be that they are benign, and a moral theory which calls that into question is highly suspect. For example, reminding agents of information relevant to a choice before they make the choice can be a nudge. Consider a sign placed above a sink in a public lavatory, reminding users to wash their hands. We may assume that at least the majority of users know that they should wash their hands, and all the reasons why it is good to do so, but they may be thinking about other things and forget. The sign is simply a reminder. Without it, less people would wash their hands. This is a nudge; the sign changed the choice architecture surrounding whether we wash our hands or not. This nudge seems extremely benign such that I believe that the position that all nudges are manipulative, at least under the trickery account, is indefensible.

Instead, we should consider if there are any sorts of nudges which are manipulative under the trickery account, and whether some forms of nudges are more susceptible to being manipulative than others.

Some nudges are not examples of agents being influenced by design of a space or system they interact with. Rather, some are simple conversations, just like other influence. One example of this given by Thaler and Sunstein is the example of framing. A doctor may need to give a potential patient information about an operation. The doctor may say something along the lines of "'Of a hundred patients who have this operation, ten are dead after five years.'" The same information can be conveyed by an alternatively phrased statement "Of a hundred patients who have this operation, ninety live after five years."[105] Thaler and Sunstein cite research that patients, even doctors, are more likely to consent to the operation if given the information framed in the latter way, than in the former.

Here we have an example of a nudge that we can easily imagine being used in a manipulative manner. This manner of phrasing clearly affects the salience of beliefs that the doctor is imparting to the patient. The doctor wants to influence the patient to believe two things. The first is that ten percent of patients die after five years, the second is that ninety percent live after five years. By framing these beliefs in the latter way, the belief about surviving is made more salient, more relevant to the patient. Presumably then, we can question whether this belief *should* be made more salient, and whether the doctor intended to make the patient more ideal by increasing the salience of the latter belief. Or alternatively, whether the doctor intended to make the patient fall short of the ideal for belief if they think the latter believe shouldn't be more salient in the mind of the patient.

This sort of nudge is what we could call a salience nudge. Noggle himself applies the trickery account to these sorts of nudges in his paper '*Manipulation, salience, and nudges*'. In this paper, Noggle examines salience nudges as a subset of nudges and how they may be manipulative. Salience is the

[104] (Sunstein C. , 2015, p. 448)
[105] (Thaler & Sunstein, 2009, p. 39)

primary way some nudges operate, as we have seen with framing, and plays a supporting role in others.[106]

Noggle argues that the trickery account applies to these nudges resulting in the following thesis:

"A salience nudge is not manipulative if it influences choice by bringing the salience of some fact into closer alignment with its actual importance. A salience nudge is manipulative if it influences choice without making the salience of any fact correspond more accurately with its actual importance."[107]

Noggle applies this thesis to the cafeteria example. Say that Carolyn designs one particular cafeteria such that the salad option is made more salient than the cheesecake option. Perhaps this is done by positioning the salad at eye level, whereas the cheesecakes higher than most people's eyeline. This would be a nudge. Noggle states that this:

"Simply highlights one of the options, without drawing attention to any fact about it that is true and relevant for the customer's decision about what to eat. The salience of the salads has been increased, but this increase does not correspond to anything that makes them more choice-worthy. In effect, it is the mere availability of the salad that has its salience heightened. … And as we have seen, influencing choice by increasing the salience of a fact not relevant to the choice at hand is a form of trickery that renders the nudge manipulative."[108]

To be clear, salience of a choice refers to the relevance ideal of belief but the salience ideal of emotion is also a method by which some nudges can function. If a choice is more salient, beliefs about it are more present, and we are more aware of it as a feature of our experience.

Noggle argues that it is only our belief that the food is available to us that is made more salient to us by changes in its position. There is nothing about the belief of availability which justifies the belief being more salient over equal beliefs regarding the availability of other foods on offer. Therefore, intentionally placing the food item in a manner which the influencer intends will increase the salience of the item's availability. This is to intentionally make the target of the influence less ideal, and thus manipulate them.

One defence Thaler and Sunstein have is that ultimately, the food items *have* to be arranged in some way. We will be influencing the customers regardless of our choice. However, in their paper '*Nudging and the Manipulation of Choice'* Pelle Hansen and Andreas Jespersen argue that this logic is akin to stating that "Because everyone must eventually die someday, one is justified in taking another's life."[109] They contend that nudges must be understood to be intentional influence in contrast of incidental influence, the latter of which only takes place as a necessary result of intentional action. The designers of the cafeteria want to efficiently serve food. They may have some awareness that incidentally some food will be more salient than other food, but they do not intend to influence the users of the cafeteria through food placement to pick certain foods. They do influence through food placement, but do not intentionally do so, only as a necessary consequence of another intentional action, serving food. This is in contrast to if Carolyn intentionally tells cafeteria staff to put low-calorie food where it will be most salient for paternalistic reasons. One is purposeful, the other incidental.

[106] (Noggle, 2018, p. 168)
[107] (Noggle, 2018, p. 168)
[108] (Noggle, 2018, p. 168)
[109] (Hansen & Jespersen, 2013, p. 10)

Another example of this distinction in practise would be the difference between a beautiful woman who owns and operates a food cart, and her designing and placing advertisements on the side of the food cart. Both will play a part in influencing passers-by to stop and buy food at the food cart. We can assume some customers are influenced by the eye-catching advertisements and what they communicate, leading them to buy something. Others will notice the beautiful vendor and stop to speak to her because they want an excuse to speak to somebody who is so attractive, and perhaps try their luck at romancing her. The advertisements are intentional influence, clearly. However, the vendor cannot help her beauty. Even if we suppose that she does intentionally and purposefully make herself look professional and approachable through her appearance, as any good salesperson would, we could equally suppose that she does this for other reasons. Perhaps, as most do, she takes care of her appearance for her own self-image, not in order to attract customers. Certainly, she does not need to *intend* to influence people through her appearance in the same manner that she might intend to influence through her appearance if she were preparing to attend a romantic dinner. If we imagine this vendor simply walking through the street, appearing beautiful and therefore influencing others, the distinction should be even more clear. For the most part, influencing others through one's appearance is incidental rather than intentional influence performed for the sake of influencing. Of course, we can imagine a litany of examples where we *do* alter our appearance for the sake of intentional influence. Yet to say that we are always doing so is incorrect. This is the crucial distinction.[110]

This line of reasoning allows us to make two useful conclusions. The first is that we should not worry about acting immorally simply by having cafeterias because we cannot help but influence people by how we arrange choice architecture. The second is that intentional interventions in choice architecture, motivated by a desire to influence, cannot claim that simply because choice architecture always exists, they are not making an active, intentional decision to influence other agents which they *are* culpable for. We do not need to worry about merely incidental alterations to choice architecture, done intentionally and with awareness they will influence, but simultaneously not being done for the sake of influencing, being considered nudges and/or intentional acts of influence. Therefore, we do not need to worry about these being considered manipulative.[111]

This is all in favour of Noggle's conclusion. However, I believe the conclusion can be extrapolated to examples which demonstrate that it can lead to some absurd further conclusions. Consider the practice of designing packaging products so that they are eye-catching and attractive to the customers. The purpose of such designs, and the intention of those who create them, is to influence

[110] This does not conflict with my previous analysis of intentional influence which involves agents as targets that the manipulator would not otherwise have targeted. This is because in those examples, the goal of the influencer is to influence a set of people, for the sake of influencing them. They were just mistaken as to the exact members of the set. They still intended to influence as a goal, rather than intentionally performing an act which they know will influence others, but not for the sake of influencing them, which would be incidental.

[111] There is a similar concern that this could be taken to contrast with the approach of manipulation as bullshit, where I asserted that we have a positive duty to avoid making others less ideal as a result of our intentional influence. Consider however, if I were to tell a sarcastic comment to a friend, knowing that other people in earshot around us might hear my comment and, without context, not realise I was being sarcastic and be misled. I do not intend to influence these people at all, so any influence I may acknowledge that I had on them as a result of my intentional action, nevertheless was not intentional, only incidental. This is a nuanced point, but I believe the distinction clearly must exist. Where it falls is a matter for further debate. I am not trying to state that a person who tells another "Run! There's a fire" in a crowded cinema is not responsible for the ensuing stampede merely because they intended for it to be a joke to a friend. Especially if they knew they would be heard and the likely consequences.

potential customers to notice and hopefully purchase one company's product over another, despite, we can assume, the products being of generally equal quality.

You may have had the experience, as I have, of shopping for a product like razorblades in a supermarket. In my experience there are countless product options in a single aisle, or sometimes a single shelf, all packaged in packaging designed to attract my attention first and foremost. After my attention has been focused on a particular eye-catching product, the packaging can proceed to communicate to me whatever the designers want to communicate to influence me to buy it.

We can assume then that at least part of the design of packaging is an intentional attempt to increase the salience of customer's beliefs about the availability of that product. Moreover, we can imagine that for at least some customers, this belief's salience is what ultimately leads them to purchase the product.

If we follow Noggle's argument, we must conclude that designing eye-catching packaging is manipulative, as it is an example of intentional influence which intends for agents to fall short of the ideal for relevance. It does so by making more relevant (salient) the belief agents hold about the availability of one product over others, in a manner not rationally justified and which the manipulative agent knows is not justified. We cannot say this is merely incidental. It clearly is intentional. I believe it would be absurd to conclude that anything other than designing packaging so as *not* to be eye-catching should be considered manipulative. It is unintuitive, both for defining it as manipulative as well as the categorisation of the act as unethical that follows from defining it so.

How then, to defend the trickery account? The solution is to deny the thrust of Noggle's argument. We can assert that mere increase in the salience of our belief about the availability of a product in comparison to others *is* a relevant reason to choose it over others.

Noggle argues that there is no reason why belief about the availability of one item of food in a cafeteria should be more relevant than any other belief about the availability of a food option. I would agree the influence does not make a belief more salient which is more relevant. However, while the belief is not relevant, paying more attention to it than other beliefs does *not impair* the chooser either.

We must consider, as we did when considering the **Surprise Party** example earlier, what the purpose of the ideal is. Except in this case, it is the ideal of relevance, or salience of beliefs, we are interrogating, rather than the ideal of true beliefs. I would suggest that the reason we want to only hold salient those relevant beliefs to the context at hand is simply due to the fact that irrelevant beliefs distract from the good or provide useless information. These 'red herring' beliefs take up cognitive resources. They are distractions, more salient over other, more important, beliefs. Increased salience of the availability of one item over another may not provide a rational reason to choose that item over others, all things considered, but it equally does not impair the chooser by taking up cognitive space which is used to attend to other 'choice-worthy' features of the items from which they can choose. It is a belief held alongside others, having little effect on the believer while not being strictly relevant itself.

For example, the fact that a product has red packaging may not be strictly relevant to the decision of whether to purchase that item over others. I will after all likely throw the packaging away as soon as I come to use the product. Yet my belief that the product's packaging is red does not take up my attention so much that it distracts me. Simply being eye-catching is not enough to impair me.

To return to the cafeteria example, Carolyn may well place the salad in more salient middle shelf, and the cheesecake further along in the corner of the cafeteria in order to nudge her students to pick what she regards as healthier options. Yet for those who absent-mindedly chose the salad over the cheesecake for no other reason than one was more saliently available to them, they were not impaired by this salience. No relevant belief was overshadowed by the salient belief about availability. Rather they clearly were not attending to relevant beliefs in the first place. Ergo, they were not made less, or more ideal by the increase in salience.

Consequently, I conclude that Noggle is wrong. I should emphasise that this does not mean that *all* salience-based nudges are not manipulative, no more than Noggle being correct would mean they *all* were manipulative. Diminishing the salience of an item, for example by hiding it away, could make an agent less ideal. A classic example is the small print of a clause in a contract. Smaller text is more easily missed or overlooked by the reader. If the small print contains content which would be relevant to the decision to sign the contract, then purposefully reducing the salience of this content would be manipulative. Rather, with nudges more so than other phenomena, there is a concern that accounts of manipulation like the trickery account will overextend and conclude that intuitively acceptable methods of influence often used in commercial environments will all be rendered equally manipulative. I believe that I have avoided this concern here.

In summary then, nudges are examples of intentional influence. Some nudges take place through the design of a system, such as a cafeteria. Others can take place in the context of a conversation and are explicitly communicated. Some nudges are the sort of influence subject to the trickery account, while others are changes to choice architecture which may be best dealt with by other theories, such as theories or coercion or manipulation as pressure. Regardless, it seems that the theory applies to them in no specific, special way. Rather, the manipulation as trickery account can be usefully applied to at least some examples of nudges without demanding we make any unintuitive or obviously wrong conclusions. Crucially, if we adopt the manipulation as trickery account, as I believe we should, then the manipulation objection to nudges, writ large, should not be a concern for advocates of nudges in public policy. Of course, other obstacles exist for those advocates outside of the purview of this project.

Dark Patterns and Dark Systems: Applying the Trickery Account

The term ‘dark pattern’[112] was coined by UX (User Experience) specialist and cognitive scientist Harry Brignull to describe a type of “bad design pattern” which is “crafted with great attention to detail, and a solid understanding of human psychology, to trick users into do [sic] things they wouldn’t otherwise have done”.[113]

[112] I note that since this section was written, the website and organisation have rebranded and now refer to ‘dark patterns’ as ‘deceptive patterns’ first and foremost. Whilst this may be more descriptive of the majority of the examples listed, other specifically manipulative design elements do not necessarily have to be deceptive, as I will expand on further in this section. This rebranding is due to “a commitment to avoiding language that might inadvertently carry negative associations or reinforce harmful stereotypes”. (Brignull, Leiser, Santos, & Doshi, 2023) Whilst understandable for that goal, it does imply incorrectly that all sorts of deceptive patterns are instances of deception. Consequently, I will maintain use of the term ‘dark pattern’ in this section to avoid what may appear to be apparent contradictions in phrasing.

[113] (Brignull, 2010)

As a UX expert, Brignull is referring to elements of apps, websites, and other user experience interfaces. However, that is not to say that the techniques of dark patterns are limited to digital interfaces. Brignull's website gives examples of types of "types of deceptive pattern".[114]

Some of these examples are straightforwardly deception, so manipulative. For example, "disguised ads" are advertisements disguised as other elements of UX design. A common example would be an advertisement placed on a page where the user downloads a product or other digital item where the advertisement itself is design to look like a "download" button. Similarly, some UX choice will misdirect users. One example listed by Brignull is for an online service where, when signing up for the service and informing the provider of the users' communication preferences, one option reads "I don't want to receive any SPAM emails, push notifications or SMS. Treatwell will only send you exclusive deals, inspiration and relevant marketing. You can opt-out of receiving these at any time in your account settings or directly from the marketing notification".[115] This is an example of misdirection because, if reading the options quickly, the user is likely to be under the impression that selecting this option will mean they do not consent to receive marketing communications. Of course, it is actually the opposite.

These examples are of communication. They operate no differently to a sign in the bathroom which reminds users to wash their hands, a news article or a face-to-face conversation. They are methods of communication, which may or may not be manipulative depending on the intent of the designers or those who employ and instruct them, as discussed in the previous chapter. They also may or may not be nudges.

What we see from both these sorts of interface design decisions as well as nudges, is that elements of a system can apparently be manipulative without necessarily communicating to agents. Rather, as the agents are motivated to engage with the system, they are influenced through this engagement with elements of the system itself.

At the beginning of this project, I considered a hypothetical example which could be considered manipulative intuitively. I will restate it here:

Cold Bureaucracy: A government wants to reduce the amount of money it spends on welfare payments for its poor citizens, without receiving the bad publicity they would if they stopped providing some payments, or directly excluded some citizens from receiving them. Therefore, they make it complicated and difficult for people to apply for the payments, by asking for documents which most poor persons do not have and cannot afford. The process is so complicated that many fewer citizens receive the payments.

An earlier study looking at the effect of web design on users listed and named some techniques which could be at play in a system such as that described in **Cold Bureaucracy**.

"*Coercion* – Threatening or mandating the user's compliance.
Confusion – Asking the user questions or providing information that they do not understand.
Distraction – Attracting the user's attention away from their current task by exploiting perception, particularly pre-attentive processing.
Exploiting Errors – Taking advantage of user errors to facilitate the interface designer's goals.
Forced Work – Deliberately increasing work for the user.
Interruption – Interrupting the user's task flow.

[114] (Brignull, Leiser, Santos, & Doshi, 2023)
[115] (Cefela, R [@riccardomc], 2022)

Manipulating Navigation – Creating information architectures and navigation mechanisms that guide the user toward interface designer task accomplishment.
Obfuscation – Hiding desired information and interface elements.
Restricting Functionality – Limiting or omitting controls that would facilitate user task accomplishment.
Shock – Presenting disturbing content to the user.
Trick – Misleading the user or other attempts at deception." [116]

Arguably all of these are taxonomies that can apply to different, hostile features of a system, and some of them do not necessarily involve communication.

Consider exploiting errors. Perhaps the benefits system's online interface is a series of forms on different webpages. If the user accidently navigates away from the page, by pressing the backspace key when not focused on a form for example, the data they entered is not saved. This may not only frustrate the user, but may lead to a less accurate application, as they may re-enter the information quicker and with less care. Another example could be a short "time-out" timer where the user is logged out of the webpage after a set period of inactivity which could be made far shorter than reasonably necessary. This, accompanied by requiring users to consult other windows, such as application guidance documents, quickly could mean that users frequently lose work.

"Forced work" is relatively self-explanatory. The longer and more convoluted a process, the less likely it is that any users other than the most dedicated will persist in performing the tasks to receive a reward. If the benefit in the example is relatively small for a user, a lengthy, complicated application process which requires the user to expend a great deal of effort will likely mean fewer users apply for the benefit.

It is sufficient I think to conclude that the trickery account can be applied to systems and interfaces designed in ways which can influence user's beliefs, desires, and emotions. However, it may be that other methods of analysing influence, such as coercion or manipulation as pressure, are best placed to explain most of them. Certainly, "forced work" appears to be an example of changing the choice situation of the user by increasing demands on the user if they choose to apply. However, we should not discount ways in which the trickery account can come into play. Many systems rely on human interaction, communication, and thus social influence, to operate. It is in these elements of systems where manipulation as trickery is most likely to be.

Algorithms and Social Media

The way social media platforms operate produces a variety of situations where an account of manipulation is useful. For example, I could address the problem of misinformation and the role of social media. This is especially relevant regarding political manipulation. Social media platforms may also be accused of acting manipulatively considering the design of some of their features and UI interfaces. However, this may largely be a regurgitation of previous segments of the account on potentially problematic nudges, and Noggle himself has considered political manipulation relatively recently which I mostly agree with.[117] Instead, it may be a better use of time to consider how social media platforms operate at their core elements. I am speaking of how social media platforms deliver content to their users using complex algorithms they usually do not release to the public. Such algorithms, and the content they provide, especially to children using the platforms, have been

[116] (Conti & Sobiesk, 2010, p. 273)
[117] (Noggle, Manipulation in Politics, 2021)

accused of influencing people in extremely negative ways. Such that we may accuse algorithms of being part of manipulative action.

A particular example is that of Molly Russell. Molly was a British teenager who suffered from depression. She was fourteen when she ended her own life in November 2017.[118] Before her death, Molly was "viewing suicide and self-harm content online"[119] on various forms of social media which use complex algorithms to present personalised content to users. In an inquest into her death, the Coroner's Court heard that she had "engaged with tens of thousands of social media posts in the six months before she died, including content which "raised concerns".[120] The Coroner in her case, H.M Coroner Mr Andrew Walker, stated in a Regulation 28 Report to Prevent Future Deaths that:

"The platform operated in such a way using algorithms as to result, in some circumstances, of binge periods of images, video clips and text some of which were selected and provided without Molly requesting them. These binge periods, if involving this content are likely to have had a negative effect on Molly. Some of this content romanticised acts of self-harm by young people on themselves. … In some cases, the content was particularly graphic, tending to portray self-harm and suicide as an inevitable consequence of a condition that could not be recovered from … The sites normalised her condition focusing on a limited and irrational view without any counterbalance of normality."[121]

This is an example of the potential power of social media. While in a lot of cases the influence it has on our lives is subtle, the purpose of social media is to influence, or at least disseminate influence. In this case, the coroner concluded that algorithmically provided content influenced a young girl to commit suicide. This has prompted the United Kingdom to consider implementing laws which govern social media websites.

Social media websites are, to an extent, in the same position as Carolyn was when deciding how to structure the cafeterias within her school district. Content they serve must be provided to its users in *some* way. The scale is also magnified. Social media platforms serve millions of people daily, all with their own personal interests, needs and wants. This is a task which requires automation in some fashion, and a social media website which cannot serve personalised content to users is the current market is not a good social media website. In my view, automation through content-delivery algorithms is not something which can be banned or removed. To suggest so it to suggest eliminating all social media platforms as we know them altogether.

However, we know that the fact the social media company must provide content in a particular way, and thus selectively present them with different influences, does not mean that the company is not responsible when it actively chooses the method if provides content.

The social media platform is also in a similar position to the editor from the previous chapter's example of **Spiteful Review.** The editor is not producing influential content themselves, they are publishing and disseminating it. However, as established in my analysis of that example, if the publisher of influence knows that the influence was created to manipulate, and the publisher publishes it anyway, they are also acting manipulatively.

I am not qualified to speak in detail about the mechanisms of how content is delivered to users; the practical challenges of doing so, at least no more than I already have mentioned; or the broader methods of creating and maintaining such algorithms, especially with the developments made

[118] (British Broadcasting Corporation, 2022)
[119] (British Broadcasting Corporation, 2022)
[120] (Lloyd, 2022)
[121] (Walker, 2022, p. 2)

recently in regard to AI. What is clear however is that ultimately an algorithm will be designed in order to produce certain result. The social media platforms will then be able to analyse how well an algorithm is able to produce the result, which the company will then use to alter the algorithm, improve it, or replace it entirely. The knowledge we need to evaluate whether an algorithm is manipulative then is not how an algorithm functions but what goal it is trying to achieve.

It is reasonable to assume that the specific goals of an algorithm, or multiple algorithms working in tandem, will differ depending on the needs of each social media platform, as each serve content in different ways and works with various different sorts of content. For our purposes, we can look at the stated goals of the platform "Facebook" in Meta's latest press release (the platform's parent company) explaining, albeit only partially, the function of their algorithm to serve users with content on their 'timeline'.

Whilst the press release explains the processes in detail, the statement of the goal of the algorithm, as far as I can tell is this. "Put simply, the system determines which posts show up in your News Feed, and in what order, by predicting what you're most likely to be interested in or engage with." Furthermore, they state: "Mathematically, things get more complex when we need to optimize for multiple objectives that all add up to our primary objective: creating the most long-term value for people by showing them content that is meaningful and relevant to them."[122]

The goal then, of the algorithm is to provide users with personalised content which is 'meaningful and relevant'. It does so by weighting posts based on the user's engagement, however that may manifest, with content they have previously been provided with or sought out on their own. It will also examine how other people with similar histories of engagement have acted. Facebook assesses 'meaningfulness' by directly surveying users.

The primary ethical concern with this algorithm is that Facebook's only method of measuring and assessing posts for any sort of quality, is engagement and user reported 'meaningfulness'. Of course, extremely misleading, and manipulative content can be very engaging. An article which lies about the actions of a politician, and intends to make readers inappropriately angry, may be considered a very engaging article by the algorithm if I leave a comment, an emoji reaction, share it to my friends and spend a long time reading it. Facebook is not concerned, at least primarily, with *why* we have engaged with a piece of influence it has served to us, only that it is engaging and we, or people like us, report is as meaningful.[123] If we compare Facebook's algorithm to the example of the editor, the algorithm is an editor who does not concern themselves with the content of the material they are publishing, only whether or not it will be read by readers and receive feedback from readers that is positive. Facebook has financial incentives to design the algorithm in this way. Users are more likely to view material they find engaging and relevant, which Facebook measures by the user's material engagement with the post in the sense of interacting with it via various metrics. They are therefore more likely to use Facebook as a platform for longer periods of time if they find Facebook provides them with engaging content. If they use it for a longer period of time, they will view more advertisements, which is one of the methods Facebook derives an income from users using the platform.

By not concerning themselves with the content, yet publishing it on their platform and essentially re-publishing it by disseminating the material to users who they think will engage with it, Facebook are

[122] (Lada, Wang, & Yan, 2021)
[123] Of course, I should note here that Facebook's algorithm is simplified to the extreme here. However, it is their simplification, not mine.

essentially in the position of either a) knowingly publishing influence they believe has manipulative intent, or b) uncaringly publishing influence where Facebook does not care if the material is intended to make users fall short of an ideal. Therefore, we have an argument that social media platforms which do not deliver content based on an algorithm which *does* show some attention to the content of the influence it publishes, are acting manipulatively. Either they are actively publishing material they know to be manipulative, or they do not care, and are thus acting manipulatively in the sense of manipulation via bullshit.

This is one of the ways we might consider that social media platforms act manipulatively. However, there is also a parallel argument that social media platforms act manipulatively merely by hosting content. After all, in **Spiteful Review** the editor receives articles and publishes them. Social media platforms do the same, allowing users to publish content immediately. This does not necessitate the involvement of a content delivery algorithm. Therefore, before addressing further any analysis of algorithm use, we should consider whether merely publishing manipulative content should count as manipulative action if social media platforms refuse to evaluate the intent of the agent producing the content the platform then proceeds to publish.

One counterargument, and important consideration, is that Facebook and other social media platforms may take the view that they have a duty to be impartial in their evaluation of content others publish to the platform, even when republishing it. While social media platforms are private platforms, there is a sense that social media is a sort of "common digital town square". Certainly, this was the self-reported view of businessman Elon Musk when he recently bought the popular social media platform Twitter.[124] It was also the view of then CEO Jack Dorsey in his testimony to the United States House Committee on Energy and Commerce in 2018.[125] The position here is that social media platforms have a duty to preserve the free-expression of their users. While there may be exceptions for directly illegal or obviously harmful content, social media platforms may take the position that it is not their place to make a determination as to the intention behind published content on their platform so long as it is plausibly genuine. If we apply this to the case of Molly Russell, the platform may take the view that content which encourages or glamorised suicide, while the platform's decision makers may themselves think of the content as misleading, inappropriately invoking emotion, or tempting users to act on desires contrary to their rational beliefs (all ways in which we could conceive of such material fitting the definition of being manipulative), they cannot say that such content was published with manipulative intent. Indeed, we can consider that the material could be made by genuine actors, or in any case without manipulative intent.

If we accept this line of reasoning, we could imagine Facebook and other content providers not as the editor in **Spiteful Review**, but as a custodian or guardian of a something like a town noticeboard, where the understanding is that all citizens have a right to publish whatever they like to the noticeboard, barring some exceptions where the content is clearly malicious. Where there is at least a plausible claim of legitimacy, it is not the place of the custodian to question the intention of any resident who wants to place something on the noticeboard unless it breaks limited rules intended to maintain the status of the board as a safe place of expression. Call this the 'custodian' versus 'editor' distinction. The former has a duty to be impartial, so cannot be held responsible for material published on the platform. The latter takes a more active role, and content they allow to be published, or remain published, reflects their own tacit approval of it. The analogy of a social media

[124] (Musk [@elonmusk], 2022)
[125] (Dorsey, 2018)

platform as one large newspaper holds true, but it is a public newspaper where the custodian is mostly obligated to accept articles.

This defence is persuasive. It does still put obligations upon social media platforms to deter obvious manipulative influence. For example, 'spam' accounts which repeatedly publish hyperlinks to fraudulent, harmful websites which might download a computer virus to the hosts computer. These are clearly deceptive and otherwise harmful pieces of influence. However, as another example, imagine a controversial financial guru who posts what would appear to be deceptive financial advice which will benefit the guru at the expense of his audience. The platform has no obligation to remove such influence. After all, they could genuinely be giving such advice. It is not the platform's place to interrogate whether actor's on the platform of being genuine if there is the appearance of authenticity, or without extremely clear evidence. If social media platforms do have such duties, then they have a ready defence for why they may allow potentially manipulative content to be publish on their platform.

However, we may be sceptical that such a duty exists. It is an argument parallel, but not directly relevant to our aims here. At the very least, we might consider that social media platforms do intervene by removing content when it threatens their platform's goals. Social media platforms have ample monetary incentives to allow potentially manipulative agents to publish using their platform if such action generates engagement and therefore revenue. On the other hand, moderating such content could cause a public backlash for the company. It could be that the 'public square' defence is a convenient excuse for private platforms to avoid expensive and complex moderation, or potentially publicly controversial moderation decisions they would need to take. We might also consider that while the defence may work for some platforms, some platforms by virtue of how they are designed and the audiences they target may purposefully place restrictions on what sort of media is published. If these restrictions are permissible, then others may be also. In that sense some platforms may be 'public squares', but not all.

We can now return to more specific case of content delivered as a result of an algorithms, where content is personalised to maximise, as an example, 'engagement and meaningfulness'. The moral failing of manipulation as bullshit is the lack of concern an agent has for whether or not influential acts they engage in will make a person less ideal. In stating that a social media platform's algorithm is manipulative, and thus the platform acts manipulatively, we are stating they either do not care, or are actively leading influenced agents astray by design. The former is more likely, barring exceptional situations where the social media platform alters its algorithm explicitly to manipulate. The simplest way for social media platforms to avoid acting manipulatively is to alter their approach. Rather than seeking to create an algorithm which maximises engagement, it should instead aim to maximise engagement *with non-manipulative material*. We should then ask how an algorithm can evaluate this.

The greatest resource for this is the data and user input social media platforms already rely upon. It should be extremely easy for users to remove content they don't want to see more of, and this should be highly weighted in algorithms. To some extent, social media platforms already allow reporting of scams, misleading information and other material. I suggest it should become a focus of the platforms. Giving users transparent access to the factors such as subject matter which influence what they see would also allow users to meaningfully choose and consent to what content they wish to see. A user who regularly accesses content that makes them feel angry, might be prompted as to whether they actually want to see more of this content, for example. The current cause of the problem is that social media platform algorithms allow users to fall into patterns of content consumption which may not, on reflection, want to continue. However, without the ability to

dramatically alter the algorithm delivering them content, changing such patterns can be difficult to do on purpose whilst still using the platform. If social media platforms can place the power in the hands of their users and alter the algorithms they use to deliver content to minimise material that users report as being manipulative or otherwise harmful, then the algorithm's designers can no longer be accused of acting manipulatively in the sense of the editor. Otherwise, I am comfortable asserting that social media platforms which do not do so to a sufficient extent are acting manipulatively by serving personalised content to users which they do not evaluate or concern themselves with beyond what engagement it can get from a user.

Conclusion

The concept of manipulation is perhaps one of more importance now than it has ever been. Sure, during the cold war period states have utilised manipulation eagerly, but new forms of technology mean we are facing ever increasing attempts to influence us in practically every environment we occupy. The more omnipresent information technology becomes, the more we will be concerned with how and why we are influenced by other agents. Consequently, it is ever more important that we know by what methods we will be influenced, and the ethical implications of those methods.

Manipulation is a concept which has developed over time. The key train of thought has been that manipulation is influence by which an agent bypasses or subverts our rational capacities. Out of this train of thought and the associated accounts, I identified Robert Noggle's account of manipulation as trickery as the most promising account. I have further built on this account. Noggle correctly identified that it is the intent of the influence which defines whether they have acted manipulatively, and that manipulation can effect our beliefs, desires and emotions in equal measure. I have argued that we should equally consider manipulative intent not just as an explicit, goal-orientated intent to lead someone astray, but also as a general disregard for whether they are led astray or not. By doing so, I have broadened the account to encompass what may be a more common sort of manipulation. Manipulation by the self-absorbed, conceited, or ignorant. I also addressed some of the follies of the trickery account, substantiated or otherwise. I have addressed apparent counter-examples such as sincere cult-leaders, strangely acting Casanova's, and apparently over included sorts of influence like the influence of pleasant aromas on house buyers. I feel confident that the trickery account, as I have expanded it, accurately accounts for our intuitions regarding manipulative acts, as well as distinguishing manipulation from other forms of influence.

A manipulative act is always a wrong committed. I understand some may balk at this commitment. Yet I find it inescapable. I find this character part of how we understand the concept of manipulation. While various ethical accounts may make space for manipulation to be a justified wrong, whether or balance or by some other measure, I commit whole heartedly to the idea that when I manipulate somebody, I wrong them, no matter how justified I may be when one considers the whole picture. Whatever you may believe about the merits of my arguments, I will echo Marcia Baron when she says "given how much some people enjoy the power they feel in manipulating others, it may seem unwise for an account of manipulation to announce that manipulation is sometimes permissible".[126] This is especially true when it comes to the implementation of manipulative technologies and other designed systems. It appears those in that sector do not yet realise that too often their technology is designed in a way which, while it may make more money, utilises a technique which is manipulative by definition. I have also considered the ethics of those other than the manipulator themselves. What liability people who have been manipulated face, both for any acts they commit as a result of being influenced, as well as for being influenced in the first place, is increasingly relevant in a political landscape where each side accuses the other of being manipulated by 'powers that be'.

I would argue the second section of this work is the most important. While a robust, complete account of manipulation as it occurs at a small scale between few agents is of course extremely useful for navigating our personal relationships, existing accounts do not bridge the gap between this sort of scale and the large-scale manipulative acts. Fundamentally, identifying the mechanics of

[126] (Baron, The Mens Rea and Moral Status of Manipulation, 2014, p. 119)

how a form of influence defined by the intent of the influencer can explain collaborative forms of media is essential to applying the same sorts of accounts to these expanded contexts. I have laid out how I think these problems can be resolved, and their implications on some examples of manipulation. Moreover, I have argued the trickery account can be utilised in considering whether nudges, as elements of influence by design are manipulative. Usefully, it does not characterize nudges as all manipulative, or seeking to nudges others as manipulative. By referring to intent, and techniques which require certain intent, we can clearly identify which nudges may be manipulative and which are mere influence.

Manipulation is unfortunately, commonplace. Having a clear definition of it, in all its forms and in all sorts of scale, allow us to identify it and oppose it when it would be used against us. Perhaps also, we should consider our own conduct. It is all to easy to reach for manipulation when we want something from another and are fearful, rightfully, or wrongly, that they would oppose us in our goals. Often times each act is a small one. A lie here or there, a misdirection when felt necessary, a bit of an emotionally poly here and there. While small, each is still a cut, and with enough cuts even the strongest bonds between people can fall apart. We can be more trusting and transparent with one another. More respectful of one another, as our status as moral and rational agents demands. Certainly, in my own study I've looked back on friendships that have been sullied, relationships soured, with a new perspective of my actions and theirs.

A final statement then. This thesis aimed to provide a developed account of manipulation and its ethics. It aimed to take the trickery account of manipulation, improve it, and apply it to contexts hereto underserved. Manipulation is a form of influence which is trickery, and it encompasses many different sorts of action. All the better then to understand it, and treat it is a tool only to be used when necessary, now we better know its consequences.

Reference List

Archer, D. (2017, March 6). *The Danger of Manipulative Love-Bombing in a Relationship*. Retrieved from Psychology Today: https://www.psychologytoday.com/intl/blog/reading-between-the-headlines/201703/the-danger-of-manipulative-love-bombing-in-a-relationship

Barnhill, A. (2014, 8). What is Manipulation? (C. Coons, & M. Weber, Eds.) *Manipulation: Theory and Practice, 21*(4), 359-379. Retrieved from http://arxiv.org/abs/1011.1669

Baron, M. (2003). Manipulativeness. *Proceedings and Addresses of the American Philosophical Association*, *77(2)*, pp. 37-54. Retrieved from https://doi.org/10.2307/3219740

Baron, M. (2014). The Mens Rea and Moral Status of Manipulation. *Manipulation: Theory and Practice*, 98-120.

Bratman, M. E. (1993, Oct). Shared Intention. *Ethics*, 97-113. Retrieved from http://www.jstor.org/stable/2381695

Brignull, H. (2010, July 8). *Dark Patterns: dirty tricks designers use to make people do stuff*. Retrieved from 90 Percent of Everything: https://90percentofeverything.com/2010/07/08/dark-patterns-dirty-tricks-designers-use-to-make-people-do-stuff/

Brignull, H., Leiser, M., Santos, C., & Doshi, K. (2023, April 25). *Types of deceptive pattern*. Retrieved April 29, 2023, from Deceptive patterns – user interfaces designed to trick you: https://www.deceptive.design/types

British Broadcasting Corporation. (2000, March 24). *Eyewitness: Why People join Cults*. Retrieved April 28, 2023, from BBC News: http://news.bbc.co.uk/1/hi/world/africa/688317.stm

British Broadcasting Corporation. (2008, August 16). *Madagascar's dance with the dead*. Retrieved April 28, 2023, from BBC News: http://news.bbc.co.uk/1/hi/programmes/from_our_own_correspondent/7562898.stm

British Broadcasting Corporation. (2022, October 7). *Chris Williamson: Bullied at school, bored on Love Island, now a podcast star*. Retrieved from BBC News: https://www.bbc.co.uk/news/uk-62917588

British Broadcasting Corporation. (2022, October 14). *Molly Russell: Coroner's report urges social media changes*. Retrieved from BBC News: https://www.bbc.co.uk/news/uk-england-london-63254635

Cefela, R., & [@riccardomc]. (2022, August 14). Nice dark pattern @treatwellnl! That didn't work. Twitter. Retrieved from https://twitter.com/riccardomc/status/1558790286964310018

Cohen, S. (2018). Manipulation and Deception. *Australasian Journal of Philosophy, 96*(3), 483-497. Retrieved from https://doi.org/10.1080/00048402.2017.1386692

Conti, G., & Sobiesk, E. (2010). Malicious interface design: exploiting the user. *WWW '10: Proceedings of the 19th international conference on World wide web*, (pp. 271-280). Retrieved from https://doi.org/10.1145/1772690.1772719

Dorsey, J. (2018, September 5). Testimony of Jack Dorsey, Chief Executive Officer, Twitter, Inc. *Twitter: Transparency and Accountability*. United States House Committee on Energy and Commerce.

Frankfurt, H. (2005). *On bullshit.* Princeton, N.J. ; Oxford: Princeton University Press. Retrieved from https://find.shef.ac.uk/permalink/f/15enftp/44SFD_ALMA_DS51248105710001441

Gillette. (2019, January 15). 'We believe: the best men can be'. *Youtube: The Guardian*. The Guardian (republished). Retrieved April 29, 2023, from https://youtu.be/UYaY2Kb_PKI

Greenberg, A. (2020). Epistemic Responsibility and Criminal Negligence. *Criminal Law and Philosophy, 14*(1), 91-111. Retrieved from https://doi.org/10.1007/s11572-019-09507-7

Hanna, J. (2015). Libertarian Paternalism, Manipulation, and the Shaping of Preferences. *Social Theory and Practice, 41*(4), 618-643. Retrieved from http://www.jstor.org/stable/24575752

Hansen, P. G., & Jespersen, A. M. (2013, March). Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy. *European Journal of Risk Regulation, 4*(1), 3-28.

Lada, A., Wang, M., & Yan, T. (2021, January 26). *How does News Feed predict what you want to see?* Retrieved from Tech at Meta: https://tech.facebook.com/engineering/2021/1/news-feed-ranking/

Lloyd, N. (2022, September 20). *Coroner has 'not forgotten' Molly Russell's family as quest for answers delayed*. Retrieved from Evening Standard: https://www.standard.co.uk/news/crime/molly-russell-inquest-delayed-north-london-harrow-barnet-b1026949.html

Long, T. (2014). Information Manipulation and Moral Responsibility. Oxford Scholarship Online.

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

Musk, E., & [@elonmusk]. (2022, October 27). Dear Twitter Advertisers. Twitter.

Noggle, R. (1996). Manipulative Actions: A Conceptual and Moral Analysis. *American Philosophical Quarterly, 33*(1), 43-55.

Noggle, R. (2018). Manipulation, salience, and nudges. *Bioethics, 32*(3), 164-170.

Noggle, R. (2021, September 29). Manipulation in Politics. *Oxford Research Encyclopedia of Politics*. doi: https://doi.org/10.1093/acrefore/9780190228637.013.2012

Pettit, P. (2003). Grounds with Minds of Their Own. In F. F. Schmitt, *Socializing Metaphysics: The Nature of Social Reality* (pp. 167-192). Rowman & Littlefield Publishers.

Raz, M. (2014, January 18). *Looking Back: Interpreting lobotomy – the patients' stories*. Retrieved from The British Psychological Society: https://www.bps.org.uk/psychologist/looking-back-interpreting-lobotomy-patients-stories

Rozencher, S. (2013, April). The Growth Of Cause Marketing: Past, Current and Future Trends. *Journal of Business & Economics Research, 11*(4), 181-186. Retrieved from https://doi.org/10.19030/jber.v11i4.7746

Scarantino, A., & de Sousa, R. (2021). Emotion. *The Stanford Encyclopedia of Philosophy, Summer 2021*. (E. N. Zalta, Ed.) Metaphysics Research Lab, Stanford University. Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/sum2021/entries/emotion/

Schneider, S. (n.d.). The Paradox of Fiction. *The Internet Encyclopedia of Philosophy*. Retrieved April 29, 2023, from The Internet Encyclopedia of Philosophy: https://iep.utm.edu/fict-par/

Strawson, P. (2008). Freedom and Resentment. In P. F. Strawson, *Freedom and Resentment and other Essays* (pp. 1-28). Taylor & Francis Group.

Sunstein, C. (2014). *Why Nudge?: The Politics of Libertarian Paternalism.* New Haven: Yale University Press.

Sunstein, C. (2015). Fifty Shades of Manipulation. *SSRN Electronic Journal, 5349*.

Sunstein, C. (2015). The Ethics of 'Nudging'. *Yale Journal on Regulation, 32*(2), 413-450. Retrieved from https://heinonline-org.sheffield.idm.oclc.org/HOL/Page?public=true&handle=hein.journals/yjor32&div=16&start_page=413&collection=journals&set_as_cursor=1&men_tab=srchresults

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online Manipulation: Hidden Influences in a Digital World. *Georgetown Law Techonology Review, 4*, 1-45.

Thaler, R., & Sunstein, C. (2009). *Nudge: improving decisions about health, wealth and happiness* (New intern ed.). London: Penguin. Retrieved from https://find.shef.ac.uk/primo-explore/fulldisplay?docid=44SFD_ALMA_DS21206113680001441&context=L&vid=44SFD_VU2&lang=en_US&search_scope=SCOP_EVERYTHING&adaptor=Local Search Engine&tab=everything&query=any,contains,nudge&offset=0

The University of Virginia Library. (2001, June 27). *The Religious Movements Page: Maryland Cults Taskforce.* (J. Hadden, Ed.) Retrieved April 28, 2023, from Religious Movements: [http://religiousmovements.lib.virginia.edu/cultsect/mdtaskforce/loomis_testimony.htm], [https://web.archive.org/web/20010627113336/http://religiousmovements.lib.virginia.edu/cultsect/mdtaskforce/loomis_testimony.htm]

Tollefsen, D. P. (2015). *Groups as agents.* Polity Press.

Treanor, J. (2008, December 15). *Revealed: desperate final hours of the world's biggest ever financial fraud*. Retrieved from The Guardian: https://www.theguardian.com/business/2008/dec/15/madoff-fraud-wall-street-news

Walker, A. H. (2022). *Regulation 28 report to prevent future deaths.* London: Crown Copyright. Retrieved April 30, 2023, from https://www.judiciary.uk/wp-content/uploads/2022/10/Molly-Russell-Prevention-of-future-deaths-report-2022-0315_Published.pdf

Walton, K. (1990). *Mimesis as Make-Believe.* Harvard University Press.

Wood, A. (2014). Coercion, Manipulation, Exploitation. *Manipulation: Theory and Practice, 1*, 1-42.