# Sharing Is Caring:
# The Epistemic Responsibilities of Social Media Services

Master's Thesis submitted for the
Research Master's Philosophy
Utrecht University
July 8, 2021

Thijs Hendrik Ringelberg
Student Number: 5733588

Supervisor: Erik Stei
Second Reader: Daniel Cohnitz
Third Reader: Joel Anderson

# Abstract

Social media corporations intervene increasingly often on their services on epistemic grounds. This thesis aims to establish the grounds for such interventions by answering the question *what are the epistemic responsibilities of social media services?* The work takes a systems-oriented social epistemological approach, focusing on contemporary social media services (SMS): content-driven services predominantly curated by means of attention-maximising algorithmic recommender systems. Arguing that the responsibilities of SMS are dependent on the epistemic effects that these services might have, considerable space is devoted to studying these effects: new accounts are developed of the role of SMS in the spread of fake news, the radicalising effects of YouTube rabbit holes, and the promotion of echo chambers on SMS. It is concluded that if social media services are to retain their societal role, they must assume substantial epistemic responsibilities in order to ensure the epistemic beneficence of the environments they offer.

## Contents

# Introduction

It is a central feature of human life that we are epistemically dependent on one another. It is difficult to imagine life without the ability to converse with others, the ability to hear accounts of events we did not witness, the ability to learn directly the lessons for which others have had to struggle. These and similar insights form the central ideas of the burgeoning field of social epistemology, which has seen great developments in the last few decades.

Those same decades have also seen the rise of a new technology that might further facilitate this epistemic interdependence: social media services. The epistemic promise that these services hold is enormous: they could significantly smoothen communication with people who are far away from us; they could ease communication in large groups; they could promote communication with people whom we would otherwise never have met. Social media services have a chance at being part of a list of great inventions that have furthered human social-epistemic abilities: they might find themselves in the ranks of written language, the printing press, radio, telephone, and television, to name a few.

Each of the technologies on this list can be characterised as follows: a speaker sends a message to an audience by means of the technology. Yet the technologies fall into two categories. In the case of some of these technologies, like the telephone and written language, responsibility for sending the message only accrues to the speaker: we do not hold the paper industry responsible for the contents of notes left on fridges; nor do we blame telephone companies if lies are told by means of the infrastructure they maintain. In the case of other technologies, like the printing press, radio, and television, some of this epistemic responsibility also accrues to those responsible for the technology. A publishing house can be held responsible for publishing a book full of blatant lies, even if the writer of the book is not an employee of the publishing house; radio and television stations can be held responsible if their broadcasts break certain epistemic standards – like prohibitions on slander, or on spreading hatred – even if those standards were only broken by guests on these broadcasts.

This brings into focus a question: to which of these categories do social media services belong? By maintaining that social media services are neutral platforms, it has historically been argued that they are part of the former category: just like phone companies, the argument goes, social media services merely maintain a communication infrastructure. What people do with that infrastructure is not the service's responsibility.

But recent years have seen some changes in this respect. Halfway the second decade of the millennium, stories started surfacing that seemed to indicate that social media services had some particularly pernicious epistemic effects. Social media services were implicated in the propagation of fake news and the promotion of conspiracy theories, to name a few. In an article in the New York Times, video service YouTube was even called The Great Radicalizer.[1] The call for social media services to play a more proactive role in preventing these effects grew increasingly louder, and over the course of the next years many services intervened at least a little. The latest significant development in this respect is the blanket social media ban of a former United States president.[2]

The present work stands in line with these recent developments. It is based on a worry that the recent epistemic interventions of social media services lack a unifying rationale: although they address some of the more obvious symptoms of the epistemic effects of social media services, like the spread of fake news, or the promotion of conspiracy theories, they fail to address the root causes of these effects. This is largely due, I think, to the fact that it is unclear what the epistemic responsibilities of social media services are. Being a new technology, we have not yet developed an intuitive feeling for the responsibilities of these services, like we have for the responsibilities of newspapers and tv and radio stations. Lacking such intuitions, a philosophical account of the responsibilities of social media services might help further this societal debate.

The central question of this work is: *what are the epistemic responsibilities of social media services?* I address this question from the field of social epistemology, which is perfectly situated to answer it due to the distinctly social-epistemic character of social media services. The work is divided into seven chapters, distributed over two parts.

The first part, consisting of chapters one, two, and three, lays the theoretical foundation. The second part, consisting of chapters four, five, six, and seven, investigates the epistemic effects of social media services in order to formulate an account of their epistemic responsibilities.

[1] Zeynep Tufekci, "YouTube, the Great Radicalizer," *The New York Times* 10 (March 10, 2018).

[2] Michael Luca, "Social Media Bans Are Really, Actually, Shockingly Common," WIRED, January 20, 2021, https://www.wired.com/story/opinion-social-media-bans-are-really-actually-shockingly-common/.

# Part One: Theoretical Foundation

The purpose of this first part, consisting of chapters one, two, and three, is to lay the theoretical foundation on which I will proceed to build my account of the epistemic responsibilities of social media services in part two.

In **chapter one**, *Social Media, or How to Organise an Ocean of Content*, I dive into the history of social media in order to formulate a workable definition of this technology. I argue that the character of social media services has changed considerably over the course of the past twenty or so years: while at first, emphasis was placed on more *social* aspects, the focus of contemporary social media services is shifting towards the distribution of *content*. This means that contemporary social media services wield curatorial power: the power to decide who sees which content at what time. The wielding of this power is delegated to complex algorithms known as recommender systems.

In **chapter two**, *Industrial Production in the Attention Economy*, I investigate the operational logic of these recommender systems. Establishing how contemporary social media services define the notion of a good recommendation is essential if we wish to understand how these services wield their curatorial power. I argue that the operational logic of these systems must be understood in the context of the business structure of social media services, which is situated in the attention economy. Following this logic, I argue, social media services should be understood as tools which aim to extract attention.

In **chapter three**, *The Epistemic Responsibilities of Media*, I build the social-epistemological framework in which this work is situated. After situating my research in the wider field of social epistemology, I formulate an account of epistemic responsibility. Media are likely to distort the messages they mediate. Therefore I argue that the epistemic responsibilities of media depend on whether media can be characterised as transparent: do they wear the distortions they cause on their sleeves?

The second part of this work, consisting of chapters four, five, six, and seven, is dedicated to answering the question whether social media can be characterised as transparent, and what conclusions can be drawn about the epistemic responsibilities of these services from the answer to that question. Because the groundwork of part one is necessary for understanding the full relevance of these chapters, I will only provide a summary of these chapters at the start of part two, on pages 49 and 50.

# Chapter One
# Social Media, or How to Organise an Ocean of Content

Before we can establish the epistemic responsibilities of contemporary social media services, we first require a definition of those services. That is the purpose of this first chapter. By comparing various historical definitions of social media, I show how these services have developed over the course of two decades. While the focus of these services was on distinctly social aspects at first, their increased popularity has gone hand in hand with an increased focus on the distribution of content. In later chapters, I will show that this change in focus is extremely significant for the epistemic responsibilities that these services have. A focus on content necessitates advanced methods of curation: it is from this need that the widespread use of algorithmic recommender system arises. And as I argue in the second part of this thesis, the design of these systems has distinctly epistemic consequences, which entail certain epistemic responsibilities.

## 1. The emergence of social media

Part of the difficulty with providing a good definition of social media is that they are a young phenomenon, which has nevertheless seen much development since its inception.[3] What came to mind when hearing the term “social media” ten years ago might be subtly different from what is generally meant with that term today. It therefore does not suffice to simply pick a definition from the academic literature, because the timing of a definition is essential for how well it covers our contemporary concept of social media. In order to provide a good definition of what we currently understand the term “social media” to mean, it is necessary to go over the history of this phenomenon, tracing the structural developments that have affected it in its short life.

The history of social media services arguably starts some 25 years ago with the introduction of so-called social network sites. Some of the earliest social network sites are Sig Degrees.com and LiveJournal, introduced in 1997 and 1999 respectively. Both of these websites are defunct by now, but that does not go for all former social network sites: the well-established social media service Facebook started as a social network site

[3] Jonathan A. Obar and Steve Wildman, “Social Media Definition and the Governance Challenge: An Introduction to the Special Issue,” *Telecommunications Policy* 39, no. 9 (October 2015): 2, https://doi.org/10.1016/j.telpol.2015.07.014.

in 2004.[4] Writing in 2007, boyd and Ellison defined social network sites "as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system".[5] With today's knowledge, what is striking about this definition of social network sites is the focus on *people* over *content*. Although it was possible to leave comments or send messages on the sites, sometimes even including pictures or videos, these bits of content were generally directed at individuals: messages landed in a private inbox, while comments ended up on the (semi-)public profile of the receiver.[6] It becomes clear from boyd and Ellison's definition that the main purpose of these sites was to create an overview of one's network of contacts, and to make it easy to get in touch with those contacts.

The term "social network sites" seems to have developed naturally into "social media": I can find no evidence for the idea that this new term was consciously introduced to indicate a change in the structure of online social services. Indeed, the terms were being used interchangeably in the early 2010's.[7] There is something to be said, however, for making a conceptual distinction between the two, because the change in use of terminology seems to track a change in the focus of online social services from *people* toward *content*. This change can be traced back in the academic literature by comparing consecutive definitions of social media with each other. boyd and Ellison introduced their definition of social network sites, which did not refer to content at all, in 2007. Just four years later Kietzmann et al. introduced a new definition, now speaking of social media rather than social network sites. This definition takes the form of a "honeycomb" of seven factors. Five of these factors describe elements that were present – although not spelled out in as much detail – in the old idea of social network sites: presence, relationships, identity, reputation, and groups. These are all factors involved in tracking and

[4] danah m. boyd and Nicole B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication* 13, no. 1 (October 2007): 212, https://doi.org/10.1111/j.1083-6101.2007.00393.x.

[5] boyd and Ellison, 211.

[6] boyd and Ellison, 213–14.

[7] See, for example, June Ahn, "Digital Divides and Social Network Sites: Which Students Participate in Social Media?," *Journal of Educational Computing Research* 45, no. 2 (September 1, 2011): 147–63, https://doi.org/10.2190/EC.45.2.b; Natalya N. Bazarova and Yoon Hyung Choi, "Self-Disclosure in Social Media: Extending the Functional Approach to Disclosure Motivations and Characteristics on Social Network Sites," *Journal of Communication* 64, no. 4 (August 1, 2014): 635–57, https://doi.org/10.1111/jcom.12106.

maintaining networks of social contacts. But interestingly, Kietzmann et al. introduce two new factors that are solely focused on content: conversations and sharing.[8] The directedness of content, which was a marker of social network sites under boyd and Ellison's definition, is on its way out: about messages (Tweets) on Twitter for example, Kietzmann et al. write that "these messages are of an ephemeral nature, without any obligation to respond".[9] About shared content, they write that "social media consist of people who are connected by a shared object".[10] This object can be shared in different senses: it might be shared in the sense that different people have a connection to the same thing, as in the case of the early versions of Facebook which were aimed at connecting students of the same university or school;[11] or it might be shared in the sense of sharing a picture, video or other piece of content on the social media service, as is the case with YouTube and Flickr.[12]

Although content starts playing a role in the definition by Kietzmann et al., this role is apparently not yet essential in 2011: in two of the four social media services that the authors map by means of their honeycomb model, the content-oriented factors do not play a role at all.[13] This changes over the next ten years. In a definition from 2015, Obar and Wildman emphasise that "(u)ser-generated content is the lifeblood of social media".[14] And in 2020, Zuckerman and Rajendra-Nicolucci define social media as "a digital space that combines communicating or sharing media with aspects of social networking sites".[15] This youngest definition clearly employs the same conceptual difference between social

[8] Jan H. Kietzmann et al., "Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media," *Business Horizons* 54, no. 3 (May 2011): 243, https://doi.org/10.1016/j.bushor.2011.01.005.
[9] Kietzmann et al., 244.
[10] Kietzmann et al., 245.
[11] boyd and Ellison, "Social Network Sites," 218.
[12] This ambiguity of the word "sharing" is not entirely clear in the paper by Kietzmann et al. However, their main source in arguing for object-centred sociality is:
Jyri Engeström, "Why Some Social Network Services Work and Others Don't — Or: The Case for Object-Centered Sociality," *Zengestrom* (blog), April 13, 2005, http://www.zengestrom.com/blog/2005/04/why-some-social-network-services-work-and-others-dont-or-the-case-for-object-centered-sociality.html.
From this blog post, it becomes clear that "sharing" is not necessarily meant in the narrow social media sense which it has today. Rather, it refers to the idea that social connections between people tend to converge around an object in a highly abstract sense, ranging from interests to schools to locations.
[13] Kietzmann et al., "Social Media?," 248.
[14] Obar and Wildman, "Social Media Definition and the Governance Challenge," 746.
[15] Ethan Zuckerman and Chand Rajendra-Nicolucci, "Beyond Facebook Logic: Help Us Map Alternative Social Media!," October 8, 2020, https://knightcolumbia.org./content/beyond-facebook-logic-help-us-map-alternative-social-media.

network sites and social media which I am emphasising: the latter are the former, *and then something*.

## 2. Definition of contemporary social media

This brings me to formulate my own definition of social media as they exist in 2021. I borrow the structure from Zuckerman and Rajendra-Nicolucci's definition, fleshing some parts out a bit more.

> *Social media are digital services that combine 1) content-functionality with 2) social network functionality.*

This, of course, requires a definition of content-functionality.

> *Content-functionality is functionality that allows users to 1) share and 2) consume content. This content is not directed at any specific individual.*

The caveat about direction of content is introduced to distinguish social media from messaging services, but also from other, more traditional means of communication such as the telephone network or the postal service.

With the term "social network functionality", I mean to refer to boyd and Ellison's classic definition of social network sites. For the sake of completeness:

> *Social network functionality is functionality that "allows users to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system."*[16]

Of course, one might disagree with this definition. I think, however, that disagreement is likely to revolve around the question whether the definition covers *all* social media. The term is used differently by different people, and one might wish to also include messaging services like WhatsApp or Facebook Messenger, for example. I would be

[16] boyd and Ellison, "Social Network Sites," 211.

perfectly happy to accept such disagreement: if a reader wishes to reserve the term "social media" for a wider or narrower group of phenomena, I would ask that reader to mentally substitute another term whenever she encounters "social media" in this text. My more important claim with regard to this definition is, first of all, that it delineates a group of existing phenomena; second, that it does so relatively unambiguously; and third, that this group of phenomena currently includes many of the more popular social media services, such as Facebook, Instagram, YouTube and LinkedIn.

### 3. The curation problem

Note that the two functionalities that define social media – content-functionality and social network functionality – are analytically and empirically distinct. The early social network sites were examples of services that employed social network functionality without offering content-functionality, and in the last section of this chapter I will argue that we are currently witnessing a rise of services that offer content-functionality but barely any social network functionality. We might even find examples of this latter possibility closer to home still: arguably, the "letters to the editor" section in newspapers meets all the criteria for content-functionality.

But a service that offers *only* content-functionality is faced with a problem: given a relatively large number of users, more content will be produced than an individual user can consume. The combined time which all users are willing to spend on creating content is likely to far exceed the time any individual user is willing to spend on consuming it. I will call this problem the curation problem: in order for the consumption-part of content-functionality to work, some kind of curation has to be performed. "Curation" in this context means 1) deciding which content to present to the user, and 2) deciding the order in which that content is presented to the user.

Combining content-functionality with social network functionality goes some way toward solving this problem. Social network functionality allows users to earmark content which they might like to see by labelling the source of that content as interesting. In this respect, a friendship, page-like or membership of a group serves the same purpose as a newspaper membership: it tells the relevant organisation which information someone would like to receive.

Social network functionality does a great job at telling the social media service which content the user does not want to see: if the user has no social connection with the source, then the content ought not to reach his eyes. It does not, however, tell the service in which order the content ought to be presented to the user. Moreover, all the combined social connections which a given user has might still make for too much content for that individual user to consume. To solve the remainder of the curation problem, contemporary social media services tend to use algorithmic recommender systems.

### 4. Algorithms and recommender systems

The concept "algorithm" is relatively vague: the exact meaning of the term is dependent on the context in which it is uttered.[17] Robin K. Hill formulates a definition that captures algorithms at a highly formal and general level: "(a)n *algorithm* is a finite, abstract, effective, compound control structure, imperatively given, accomplishing a given purpose under given provisions."[18] An algorithm is like a highly specific list of instructions that can be carried out without requiring any interpretative creativity. The litmus test for this latter requirement is whether the instructions can be carried out by a computer.

A recommender system is an algorithm, or collection of algorithms, that has as its purpose to solve the recommendation problem. The recommendation problem can be seen as a more concrete cousin of the curation problem: while my statement of the curation problem was the mere observation that there is more content available than an individual user can consume, the recommendation problem is solved by providing an answer to the question "with which content should the user be presented next?" in a specific situation.[19] Employing information about users, content, and the way users have interacted with content in the past, the recommender system can calculate which piece of content might be good to recommend.

Some crude examples will help make clear how these systems work. Imagine a platform offering video content. A user can either watch a video to the end, which is seen as an indication that the user likes the video; or the user can stop watching early, which is seen as an indication that the user dislikes the video. The platform uses a recommender

[17] Andreas Tsamados et al., "The Ethics of Algorithms: Key Problems and Solutions," *AI & SOCIETY*, February 20, 2021, 2, https://doi.org/10.1007/s00146-021-01154-8.

[18] Robin K. Hill, "What an Algorithm Is," *Philosophy & Technology* 29, no. 1 (March 2016): 47, https://doi.org/10.1007/s13347-014-0184-5.

[19] Dietmar Jannach and Gediminas Adomavicius, "Recommendations with a Purpose," in *Proceedings of the 10th ACM Conference on Recommender Systems* (RecSys '16: Tenth ACM Conference on Recommender Systems, Boston Massachusetts USA: ACM, 2016), 7, https://doi.org/10.1145/2959100.2959186.

system which aims to recommend videos that users will like. A specific user, Alfred, has just watched entirely videos α, β, and γ. Previous data shows that many other users who watched all these three videos to the end, tended to stop short video δ, while they also watched to the end video ε. Based on this data, the recommender system recommends video ε. If Alfred watches video ε to the end, the recommendation is booked as successful.

In the preceding example, the recommender system works on the principle of "collaborative filtering". This principle assumes that agreement between people in the past is a predictor for agreement in the future: by creating a group of users who have interacted with the same videos in a similar way, the system creates a dataset which it can use to predict the behaviour of an individual user in a given situation. A great advantage of this type of recommender system is that no information is required about the content itself: in the example, the videos were not categorised in any way except by how users interacted with them. Since interpreting and understanding content to such an extent that useful categories can be made is an extremely difficult task for digital systems, a recommender system that merely uses readily available and easily scraped data is a big advantage.[20]

Another principle underlying recommender systems is content-based filtering.[21] Imagine the same video platform again, except now the videos are categorised: videos α, β, γ, and δ are all category 1, while video ε is category 2. Assuming the same behaviour on the side of Alfred as was just described, a recommender system using content-based filtering will recognise that Alfred has a preference for videos of category 1, since that is all he has so far watched. Because it is in the same category, the recommender system will recommend video δ over video ε. A clear disadvantage of this kind of recommender system is that it requires a pre-existing categorisation of content. In some cases, such a categorisation is readily available: a platform that allows text-based, photographic and video content gets those three categories for free. But in many other cases, creating accurate categories is difficult, as it either requires human intervention to classify bits of content, or sophisticated artificial intelligences to attempt the same.

It also becomes clear from these examples that the goals and measures which are used in the creation of the recommender system are highly relevant for how the system will

[20] Francesco Ricci, Lior Rokach, and Bracha Shapira, "Recommender Systems: Introduction and Challenges," in *Recommender Systems Handbook*, ed. Francesco Ricci, Lior Rokach, and Bracha Shapira (Boston, MA: Springer US, 2015), 12–13, https://doi.org/10.1007/978-1-4899-7637-6_1.
[21] Ricci, Rokach, and Shapira, 11–12.

perform. Typically, the goal of a recommender system is defined as *finding good items*,[22] but this only shifts the focus: how should “good items” be defined? In the examples I just provided, the definition of a good video was a video which the user liked. This gets us a bit further, but it is still clearly ambiguous. To understand what the recommender system optimises for, we must consider the definition of a liked video as it was operationalised in the recommender system: a liked video is a video that the user watched till the end. The *Operationalised Definition of Success*, then, is to cause the user to watch videos till the end.

Of course, the examples of recommender systems I have provided are extremely simple. More kinds of recommender systems exist, and all these different kinds of systems can be combined into even more complex systems. A social media service which aims to solve the curation problem by means of a recommender system is likely to employ an extremely complex recommender system: not only are the amounts of content enormous, and enormously varied; there are also incredibly many users producing diverse kinds of data. Incorporating all the data which is available to such a service into a single recommender system is a complex task with a complex solution. But although the operational structure of the recommender systems might be much more complex, we can still expect the teleological structure to be broadly the same. Being an algorithm – or collection of algorithms, integrated into a single algorithm – recommender systems need to function without requiring any interpretative creativity. The definition of algorithms also stipulates that algorithms are finite. It therefore needs to be clear when the recommender system has accomplished its task. These two facts combined mean that there must always be some operationalised definition of success. Such a definition might increase in complexity in the sense that it weighs different requirements – the goal might be to cause two thirds of the videos to be watched to the end, and one third to be stopped prematurely, for example – but the definition must be clear, for if it is not, then the algorithm cannot work properly. I will further take up this notion of Operationalised Definitions of Success in chapter two.

[22] Jannach and Adomavicius, “Recommendations with a Purpose,” 7.

## 5. A futuristic definition of social media

So far, I have shown that modern social media services incorporate social network functionality *and* content-functionality, using the former to simplify the curation problem which is inherent to the latter. But social network functionality is not enough to entirely solve the curation problem: to solve the last bits of it, contemporary social media services use recommender systems.

Now, I ask you to reconsider the history of social media definitions which I sketched in section 1. There I argued that social networks slowly incorporated content-functionality, until they came to be the social media services which we know today: services with one foot in social network country, and one foot in content land. But I believe that there is more to this development that meets the eye: it does not merely show a gradual incorporation of content-functionality, but it shows that content-functionality is slowly usurping the position that social network functionality once had. Over the years, social media seem to have turned from places you visit to meet and track your friends, into places you visit to consume content. If this analysis is correct, then a simple extrapolation toward the future tells us that the dual functionality structure that social media have today – content-functionality combined with social network functionality – is merely a half-way point in a larger development of social media from social networks to content services.

This "analysis" would be little more than speculation, if the first signs of the endpoint of this development were not already in sight. According to the analysis, we would expect a fully "developed" social media service to have ditched its social network aspects, leaving the solution of the curation problem entirely to a recommender system. But such a creature roams the earth already: it is an app for sharing short videos, and its name is TikTok.

Admittedly, TikTok has not entirely let go of its social network functionality: it is still possible to follow your friends, and to share things with them. However, the service works perfectly without following anyone. In the words of John Herrman from the New York Times:

> *(…) the first thing you see isn't a feed of your friends, but a page called "For You." It's an algorithmic feed based on videos you've interacted with, or even just watched. It never runs out of material. It is not, unless you train it to be, full of people you know, or things*

> *you've explicitly told it you want to see. It's full of things that you seem to have demonstrated you want to watch, no matter what you actually say you want to watch.*[23]

Tiktok is not, I think, a strange little brother of "conventional" social media platforms: it is the logical next step in the development of social media. Earmarking content as interesting by means of social network functionality may have once been a useful solution to a big problem; in the age of increasingly powerful artificial intelligences it is a crude tool that throws out if not all of the baby then at least the majority of it with the bathwater. Why obscure the overwhelming majority of content from your users merely because they do not know its source? TikTok demonstrates that the curation problem can be solved entirely by means of recommender systems, and is hugely successful because of it.

Being on the frontier of social media development, TikTok challenges my definition. I therefore propose a new definition of social media:

> *Social media are online services offering content-functionality, combined with some form of automated and personalised curation.*

The vagueness in the specification of curation allows this definition to cover both social media services that employ social network functionality and those that do not. As such, I believe this definition to not only be more future-proof than my previous one, but also more general.

## 6. Implications for social epistemology: the project before us

Social epistemological studies of social media tend to focus on the social aspects of these services: much has been written about the epistemic effects of social network structures.[24] This is understandable, perhaps: after all, it is the social focus that social epistemology shares with social media. But I have argued that the social aspect of social media is waning. Social media, once purely *social*, might soon be purely *media*, merely focused on distributing content. Such new social media will require new social epistemological conceptualisations. It is this project that I intend to start in this work.

[23] John Herrman, "How TikTok Is Rewriting the World," *The New York Times*, March 10, 2019, https://www.nytimes.com/2019/03/10/style/what-is-tik-tok.html.

[24] See, for example, Regina Rini, "Fake News and Partisan Epistemology," *Kennedy Institute of Ethics Journal* 27, no. 2S (2017): E-43-E-64, https://doi.org/10.1353/ken.2017.0025.

The move from social network based social media services to content-based social media services goes hand in hand with an increasingly active role that is afforded to the service itself. Social network sites were characterised by directed content: the originator of the content decided who it was sent to. The *un*directedness of content on contemporary social media means that the service itself has the power to decide who gets to see which content at what time: I will call this power a service's *curatorial power*. The next two chapters aim to characterise this curatorial power of social media services. In chapter two, I dive deeper into the workings of social media recommender systems. The central question there is what the Operationalised Definition of Success of contemporary instances of such systems might look like: what are the criteria on the basis of which contemporary social media services curate their content? In chapter three, I provide an account of the epistemic responsibilities of media in general, which will help us understand how epistemic responsibilities might flow from the curatorial power of social media services.

# Chapter Two
# Industrial Production in the Attention Economy

In the previous chapter, we saw that social media increasingly curate their content by means of recommender systems: (collections of) algorithms geared toward providing an answer to the question "which content should the user be presented with next?" Recommender systems need some specification of what they are supposed to do: for what kind of goal are they supposed to strive? Such a goal can be expressed on multiple levels of abstraction. For example, a programmer might want the recommender system to recommend videos which the user will like, but "liking" is underdefined. In order to implement this goal into the algorithm, the programmer will have to choose a definition of "liking" which a computer can work with. One example of a definition of a liked video is 'a video which the user watches till the end'; another example might be 'a video which the user shares on his personal page'. I have termed such a definition which a computer can work with an *Operationalised Definition of Success (ODS)*.

In most cases, the ODS is chosen at the discretion of the social media service itself. Thus a social media service's choice of ODS is an expression of how that service wields its curatorial power: it is by selecting a specific ODS that the service decides who gets to see which content at what time. The goal of this chapter is to establish exactly how contemporary social media services wield their curatorial power: what might the ODS of contemporary social media recommender systems look like?

I start with a discussion of the apparent neutrality of recommender systems. Then, I discuss how trade secrets form a barrier to this research. Because the source code of contemporary social media recommender systems is not in the public domain, I am forced to deduce the ODS of contemporary social media recommender systems from an analysis of the business logic of social media services – which I do in section three. This leads me to the topic of attention maximisation, which I further address in section four.

## 1. The apparent neutrality of recommender systems

Superficially, recommender systems have an air of fairness. Curation of content is sensitive work: it has the effect of deciding what an individual user gets to see and can therefore be viewed as a form of censorship. We might be doubtful about letting a human perform such work if we are unwilling to be influenced by the biases – or even the bad intentions – of our curator. A computer, on the other hand, seems a neutral arbiter: computers do not vote, cannot fall for ideologies, and – most importantly – are notoriously bad at understanding human language. Recommender systems, then, must have a certain blindness for the content they work with: they recommend things, but do not have the capacity to understand what they are recommending. Collaborative filtering, which was explained in section 4 of the previous chapter, is a great example of this blindness: the only input which this recommender system technology uses is data about the way other users have interacted with a given piece of content. Like Lady Justice, a recommender system based on collaborative filtering seems to render its judgements whilst wearing a blindfold.

But the way we interact with something is partially decided by that thing itself. I interact differently with a poster advertising a concert than with a political pamphlet. If people structurally interact with a certain kind of content in a certain way, then a recommender system could develop a bias with respect to that content. Let's revisit my pet example: the ODS of "video watched till the end" as a proxy for "liked video". Some videos I might not watch till the end *because* I like them. Think of a video posing a riddle, providing the answer at the end: such a video I will consciously turn off early in order to think about the riddle myself. It will only be at a later time that I reopen the video in order to watch the last minute. A video like this is unlikely to perform well – that is, be recommended a lot – under the ODS in the example. Conversely, some other kinds of videos might be expected to perform extra well under the ODS, even though those videos are not necessarily more "liked".

These are not world-shaking problems. It is hardly surprising that a system will have some bias for certain kinds of content. And a believer in the fairness of recommender systems might say: would we not much prefer these kinds of silly biases for content with certain structural features over the really malicious kinds of biases that human curators can have? Would it not be much more harmful to be exposed to curation that might be *epistemically* biased – by which I mean leading us to a certain conclusion, in the sense that a state-issued censor might curate the newspaper in such a way that it leads citizens to the conclusion that the Supreme Leader is great – over curation that is, at most, *structurally* biased? And does the blindness that recommender systems have for the content they recommend not guarantee that such epistemic biases cannot be implemented by means of them?

In consecutive chapters, I argue that the answer to this latter question should be "no". The idea is that not only structural, but also epistemic features of content can cause people to interact with certain content in a certain way. A recommender system that makes recommendations based on people's interactions with content can therefore function quite analogously to the human censor: at least in some cases, recommender systems can be epistemically biased.

But before I can get to that argument, I first need to clear the ground a little. In order to predict the kind of epistemic effects we might expect to be caused by a certain choice of ODS, I first need to settle on a specific ODS to investigate. In the next section I explain why this choice is not entirely obvious due to certain informational constraints, after which I use what evidence is available to construct the central ODS of this work.

## 2. Trade secrets are an obstacle to this research

In order to see how recommender systems can develop epistemic biases, I would ideally perform experimental research with the relevant recommender systems. I would provide the system with content and a set of users, and observe its behaviour. I would observe whether tweaking certain features – for instance, changing the ODS, changing the relevant data, or changing the behaviour of the users – might have an effect on the kind of content that is recommended. Such research, although extensive – after all, there are many different kinds of recommender systems, as well as kinds of content and users – would be relatively easy to perform, since the recommender system would be entirely under my control.

Sadly, however, such research is not possible. The inner workings of the most common recommender systems, that is, those systems which are used by large social media corporations, are well-guarded trade secrets.[25] The owners of these algorithms are unwilling to publish data about their systems such as objectives, design, and training data. Indeed, social media corporations even tend not to offer functionality for researchers to observe the behaviour of their recommender systems.[26]

[25] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi, "Recommender Systems and Their Ethical Challenges," *AI & SOCIETY* 35, no. 4 (December 2020): 958, https://doi.org/10.1007/s00146-020-00950-y.

[26] "AlgoTransparency Manifesto," AlgoTransparency, accessed March 25, 2021, https://algotransparency.org/.

Lacking such information, my second best option is an attempt at reconstructing the kind of recommendations we might expect, given what we *do* know about the structure of social media corporations as they currently exist. In the next section, I will formulate an ODS of which I will argue that it is plausible that contemporary social media services employ it, based on the business models that these social media services are a part of.

There are, of course, limitations to this type of research. A first limitation is the fact that there is no way to be sure that contemporary social media corporations do indeed use a version of the ODS which I will consider. This is not, however, of very great concern. An important goal of this work is to establish that epistemic considerations need to be taken into account when formulating the ODS, and when building the wider recommender system. By showing that there is at least one ODS which *can* be expected to have epistemic effects, the link between ODS and possible epistemic consequences has been established, which is enough to give substance to the idea that epistemic considerations ought to be on the table in the design process of the recommender system.

A second limitation is the fact that the degree of complexity and success of contemporary recommender systems is unclear. The relationship between ODS and epistemic structures holds conceptually, as I will argue in the ensuing sections, but there is no guarantee that contemporary recommender systems actually pick up on this relationship. From my theoretical vantage point, all I can do is argue that a sufficiently sensitive recommender system will pick up on the relationship which I describe. What "sufficiently sensitive" means in this context, however, is an empirical question to which I cannot provide an answer. See also chapter six for a more elaborate discussion of this point.

All in all, there are two ways in which this and consecutive chapters can be read: they can be read as a warning that the recommender systems employed by contemporary social media have certain epistemic effects, and they can be read as an argument that there is an analytical connection which holds between the design of a recommender system and the likelihood that content with certain epistemic features will be recommended. What the two limitations which I just discussed show, is that the former of these two arguments, the warning for contemporary social media services, is speculative to a certain degree. Indeed it is speculative by necessity, because the relevant information to assess the soundness of the argument is intentionally being kept out of the public domain. I believe that there are good reasons to expect that the argument is actually sound, and I will point those reasons out along the way. But even if my findings do not correspond with empirical reality, the relevance of these chapters as an argument for the analytical connection between the design of recommender systems and the likelihood that content with certain epistemic features will be recommended, remains.

## 3. Industrial production in the attention economy

An important reason for the huge popularity of today's social media services is that using them is free. This raises the question: if users are not paying customers, then how do social media corporations make their money? Mark Zuckerberg, CEO of Facebook, answered this question succinctly after he was asked it during the hearing by the United States Senate on privacy and disinformation on Facebook: "Senator, we run ads."[27]

Social media corporations generally make their money by selling advertising opportunities to other organisations. From a financial perspective, this is the *raison d'être* of these corporations. A corporation which makes its money by selling shoes exists to sell shoes; similarly, a corporation which makes its money by selling advertising opportunities exists to sell advertising opportunities. Of course, it is possible for a company to have another goal than merely to make money: a fashion brand might aim to make the fashion market carbon-neutral, a technology company might aim to further the world's computing technology. But at the bottom line, a fashion brand that fails to sell clothes will go bankrupt, and a technology company whose computers do not leave its warehouses will not have funds to spare for research. The financial motive is the only motive which cannot be erased from the bookkeeping of a corporation's motives, and for social media corporations the financial motive is to sell advertising opportunities.

The term "advertising opportunities" is a bit vague. Merely an opportunity to place an ad somewhere constitutes an advertising opportunity, but no organisation will be interested in buying that opportunity if "somewhere" means "in the middle of a desert". An organisation places an ad in order for it to be seen, and the more people who see the ad, the better. This is what makes highways such a great place for billboard placement: thousands pass by them every day. Clearly, then, what is valuable is not the advertising

[27] United States Government, "Transcript of Mark Zuckerberg's Senate Hearing," *Washington Post*, April 11, 2018, https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/.

opportunity, but the potential consumer viewing the ad. Advertising opportunities are valuable due to the fact that they generate ad views.

Thus my wish to see social media corporations through an economic lens as productive firms leads me to say that these corporations produce ad views. Ad views are a relatively uniform product. For an organisation placing an ad, it does not matter whether its audience sees the ad in a newspaper, on a billboard, or through a social media service. Social media corporations, and other internet corporations, do spend some effort attempting to differentiate their products. This is mainly done by means of "personalised advertising", the idea of which is that ad views can be made more valuable if they are better targeted at a specific audience. Although much has been written about this practice,[28] I will disregard it in what follows.

A more important question is the following: if ad views are the products which social media corporations produce, then which resources are required for this production process? We can split resources into two categories: those things which do the producing, and those things from which the product is made. Ignoring traditional economic terminology for a moment, I will call the former *productive resources*, and the latter *raw materials*. The productive resources of a social media corporation are the machines the corporation uses, i.e. their servers and the software running on those servers; and the people that created those machines, i.e. the software developers. Harder to establish is what constitutes the raw materials. Of which stuff is an ad view composed? At first glance, there seem to be two things in play: an ad, and a person viewing the ad. But the first of these two is uninteresting from the producer's perspective, because it is supplied by the customer: compare the ad to the feet of the shoe shop's customer. The only thing the *producer* – the social media corporation – supplies in this productive process is the person viewing the ad. A social media corporation creates ad views from the raw material called human attention.

Thus, when viewed through this lens a social media service is best described as an extractive tool. Like a drilling rig exists to extract oil from the earth's crust, so a social media service exists to extract attention from humans.[29]

[28] For a critical take on the effectiveness of personalised advertising, see Cory Doctorow, *How to Destroy Surveillance Capitalism* (New York, NY: Stonesong Digital, 2020).

[29] This picture of social media as an extractive tool is an important part of the rhetoric of the *Center for Humane Technology*. This organisation, which is run by various programmers who formerly worked at different digital tech corporations, "is dedicated to radically reimagining our digital infrastructure." It spreads its message by various means, including a podcast called "Your Undivided Attention" and the

These are not new ideas. What I have been describing is a business model that fits the paradigm of the attention economy. The underlying idea for the attention economy was first introduced by Nobel-prize winning economist Herbert A. Simon. In a world with an endless supply of information, Simon argues, information is no scarce resource. Therefore the scarcity description flips around. In an information-rich world, that which is scarce is the thing required to consume information: human attention.[30] Pieces of information, or information providers, compete for this attention.

Simon writes in 1971, and one of his main concerns is when to harvest information from the world. Under his interpretation, there is no scarcity of information because it is always available to be harvested all around us: if I wish, I can harvest new information right now by counting how many objects are on my desk, or by listening to the sounds I hear coming from outside. This causes *information overload*: my attention is finite, so I cannot attend to all the information that is available to me. I must *filter* information in order to attend to only the most important bits. This, Simon argues, is where an *information processing subsystem (IPS)* can help. An IPS might be a computer system, it might be a person or a group of people: important is that it is a system which receives information input and delivers information output. Think of an intelligence agency providing information to a government, or even of the oil indicator light on a car's dashboard which only lights up when the oil level gets dangerously low. An IPS can help with the information overload problem "only if it absorbs more information previously received by others than it produces – that is, if it listens and thinks more than it speaks."[31]

It is useful to compare Simon's notion of an IPS to a social media service's recommender system, as the differences between the two make clear how the structural logic of a recommender system affects the kind of output we can expect from it. According to the light definition of information processing subsystems which I just supplied, a social media service's recommender system is an IPS: it receives information input – user

recent Netflix movie "The Social Dilemma" – both are included in the reference list of this work. The work of this organisation played an important role as inspiration for my own ideas. Yet, while the Center is good at pointing out pertinent problems, I find that some of the theoretical underpinnings it provides are lacking from a social and human perspective. I make some more remarks on that topic in chapter five. Part of the aim of the present work is to provide a more firm theoretical background, grounded in the humanities, to some of the claims that the Center makes. Quotations from "Center for Humane Technology," accessed June 16, 2021, https://www.humanetech.com/.

[30] Herbert A. Simon, "Designing Organizations for an Information-Rich World," in *Computers, Communications, and the Public Interest* (Baltimore, MD: The John Hopkins Press, 1971), 40–41.

[31] Simon, 42.

content – and delivers highly personalised information output. Yet there is one important difference with Simon's story: the purpose of a social media recommender system differs from what Simon describes. According to the social media service's business logic, the social media service is a machine for extracting a user's attention. This means that, from the perspective of the service, the goal must be to show as much content to the user as possible. Of course, the recommender system could never show all content to any individual user: the sheer amount of content is simply too great to do this. But if it could, it would, for that would serve the business goal best: showing a user more content means receiving more attention from the user, and attention can be processed into ad views. This runs counter to what Simon recommends for an IPS: an intelligence agency that simply aims to absorb as much of the president's attention as possible is a bad intelligence agency. An oil indicator light which does everything in its power to seize and keep the driver's attention is a hazard.

Does this mean that the business structure of a social media service is in opposition to the user's interests? It is starting to become clear that information processing subsystems can have different operative logics, and that choosing the correct operative logic involves considering what one wants the system to do. The contemporary social media service has its operative logic dictated by its business model, the purpose of Simon's IPS has a different teleology in mind. The present work investigates the question what the operative logic of social media services ought to be from an epistemic perspective.

In conclusion, I have argued that social media corporations can be expected to aim for *attention maximisation*. The more attention the service can receive from its users, the more ad views it can produce. By means of the attention economy paradigm I have made clear how this goal fits with the kind of business a social media corporation is. But my reasons for claiming that attention maximisation must be one of the goals of social media services are not merely deductive. Although it concerns highly guarded trade secrets, there is empirical evidence which points in this direction as well. First of all, there are scarce remarks by social media corporations themselves. In a 2012 blog post, for example, a YouTube employee remarked that "(..) when we suggest videos, we focus on those that increase the amount of time that the viewer will spend watching videos on

YouTube, not only on the next view, but also successive views thereafter."[32] Secondly, there are ample remarks from professionals who formerly worked as developers for social media services that point in the same direction.[33] In fact, it is a widely held view that social media services perform some sort of attention maximisation. This will therefore be the ODS which I will examine in this thesis: a recommendation is successful if it manages to lengthen the time spent by the user on the service.

### 4. Strategies for attention maximisation

What does it mean to maximise attention? A first question which bobs to the surface is what attention is, exactly. Much has been written about this question, but I wish not to dive into it too deeply: a simple garden variety conceptualisation of attention is enough for my purposes. Davenport and Beck, who wrote the first book on the attention economy from a business perspective, define attention as "focused mental engagement on a particular item of information."[34] This definition will do at present: a social media service extracts attention from a user when said user focuses his or her mental engagement on a particular item of content which is being shown on the platform.

When thinking in attention-economic terms, it helps to conceptualise attention as a resource which always exists. A person's attention is like the beam of light coming from a flashlight that cannot turn off. If the beam is not illuminating *this* thing, there must be some *other* thing which is illuminated by it. This requires some stretching of our everyday concept of attention: it means that we must redescribe situations in which we would usually say we are not paying attention, as situations in which we are paying attention to something else. A day-dreamer pays attention to his daydreams, a distracted student pays attention to her phone. If we conceptualise attention in this way, then we can see the attention market as a zero-sum market: gains in attention by one player on the market must necessarily result in a loss of attention for some other player.[35] If the goal of a player on the attention market is to garner as much attention as possible, then there are two kinds of strategies which said player might use: the attention-garnering player might either *actively* try to draw attention, or it can prevent attention from being directed at its

[32] Eric Meyerson, "YouTube Now: Why We Focus on Watch Time," *Blog.Youtube* (blog), accessed March 3, 2021, https://blog.youtube/news-and-events/youtube-now-why-we-focus-on-watch-time/.

[33] For a bundle of such testimonies, see Jeff Orlowski, *The Social Dilemma*, Docudrama (Netflix, 2020).

[34] Thomas H. Davenport and John C. Beck, *The Attention Economy: Understanding the New Currency of Business* (Boston, Mass: Harvard Business School Press, 2001), 20.

[35] Davenport and Beck, 94.

competitors – I will call this *passive* attention-drawing. Due to the zero-sum nature of the market, both active and passive attention-drawing strategies result in more attention garnered by the player employing them.

Absent the ability to physically force people to pay attention to something, active attention-drawing strategies are dependent on somehow enticing a person to pay attention. Common sense informs us that there is a difference between capturing someone's attention, and holding on to it. A man suddenly shouting in the street is likely to capture the attention of most passers-by; if he keeps shouting the same thing at regular intervals, he will be sure to have lost most of that attention within a minute. On the other hand: if the man, after shouting once or twice, performs a magic show, he will likely hold on to the attention of many more passers-by. A good active attention-drawing strategy must be aware of this: it must provide an initial "hook" to draw attention, followed by something else which serves to hold on to that attention – continuing the fishing analogy, we might call this the "line".

Passive attention-drawing strategies, on the other hand, rely on keeping people from paying attention to certain things. At first sight, this might seem like an impossible task: if we include daydreams and passing cars among the things to which a person can pay attention, then a person will always have the option to pay attention to something else than me, regardless of how many possible objects of attention I eliminate. Passive attention-drawing strategies can nevertheless be effective, because people pay attention to things in order to get something from those things: Vanessa reads a newspaper to learn the latest news, Mark watches tv at 8 pm on Saturday in order to see a movie. Therefore eliminating alternative ways of getting the thing I offer is an effective way to ensure that people will come to me. If a newspaper eliminates all of its competitors, it can expect to enlarge its own customer base.

## 5. Conclusion

The aim of this chapter was to characterise the way social media services wield their curatorial power by means of the choice of ODS. Although information about social media recommender systems is scarce due to the fact that such information is regarded a trade secret, I argued that the business model of social media corporations makes it likely that the ODS of their recommender systems enshrines some kind of attention maximisation. I identified three attention-maximising strategies. Two of these are active strategies, termed the hook and the line. Third come passive strategies.

These three types of strategies will serve to structure the discussion of part two of this thesis. In chapters four, five, and six I will consider each of the three strategies in turn, considering how the implementation of these strategies might have certain epistemic effects.

Before I can do that, however, a final and important bit of groundwork is necessary. Having now characterised social media services as well as the way these services wield their curatorial power, it is still unclear how these matters relate to epistemic issues, and specifically issues of epistemic responsibility. I address these topics in the next chapter.

# Chapter Three
# The Epistemic Responsibilities of Media

The purpose of this work is to establish the epistemic responsibilities of social media services. Having concerned myself with the definition and workings of social media services, I can now turn to the epistemic side of things. The central question which this chapter aims to provide an answer to, is: what is the relationship between the activities of media and epistemic responsibility? Simultaneously, this chapter provides the epistemological context in which this work is embedded.

I start by situating the present work in the wider field of social epistemology. From this discussion flow two further discussions regarding my conceptualisation of epistemic agents, and my measure of epistemic evaluation. Having addressed these issues, I start constructing my notion of epistemic responsibility. I do this by connecting Goldberg's recent account of epistemic responsibility in belief formation with his account of assertoric responsibility. After providing an overview of Goldberg's ideas, I further elaborate on this account by means of a discussion of the epistemic power of media, which – I argue – necessitates an explication of the relationship between epistemic responsibility and media activities. I conclude with a short discussion of the relationship between epistemic responsibility and social media services.

## 1. Supply-side epistemology

In chapter one, I defined social media as online services offering content-functionality, combined with some form of automated and personalised curation. According to this definition, social media services are clearly epistemic systems in the sense that Goldman defined them: "a social system that houses social practices, procedures, institutions, and/or patterns of interpersonal influence that affect the epistemic outcomes of its members."[36] Regardless of whether social media still incorporate social network functionality, content functionality alone constitutes complex patterns of interpersonal influence by dispersing content from a single user to an audience of users. Being a study of an epistemic system, the present work is an instance of what Goldman calls "systems-

[36] Alvin I. Goldman, "A Guide to Social Epistemology," in *Social Epistemology: Essential Readings*, ed. Alvin I. Goldman and Dennis Whitcomb (Oxford ; New York: Oxford University Press, 2011), 18.

oriented social epistemology" (henceforth: SYSOR SE). But due to the peculiarities of contemporary social media services, this is only a partial characterisation.

Goldman provides many examples of systems that can be studied under the header of SYSOR SE, like science, education, journalism, and the legal trial. In most of these examples, what makes the system systematic are certain social customs, norms and expectations. A legal system, for example, is defined by the roles that participants assume: there are no practical barriers that keep the judge from assuming another role, apart from the fact that that is not how these things go. This means that an evaluative study of the legal system cannot make a clear-cut distinction between the system that is studied and the people who operate within that system: if the people operated differently, then it would be a different system. Thus in order to answer evaluative questions about the epistemic system (e.g. what is a good legal system?), it will often be necessary to answer evaluative questions about individuals who participate in the system (e.g. what should the goals of the judge be?).[37] We might call such systems *dynamic*, and the type of epistemology that studies them *dynamic SYSOR SE.*

Social media services, on the other hand, know a clear distinction between *users* of the system, and the system *itself*. The system – specifically, the methods of content curation – can be changed without changing the role that is played by users, and vice versa. It is in this sense that the system can be called static: it can be characterised independently from user behaviour. The patterns of interpersonal influence that characterise these services as social epistemic systems are largely contained in the design of the system, and are often even obscured to the users themselves. In these kinds of systems, it is therefore possible to answer evaluative questions about individuals (e.g. what is proper epistemic conduct for a social media user?) independently of evaluative questions about the system (e.g. what is an epistemically good social media service?). In a nod to economic theory, we might call social epistemological studies that aim to answer questions of the former kind instances of *demand-side*, and studies that aim to answer questions of the latter kind instances of *supply-side SYSOR SE*. Demand- and supply-side SYSOR SE can function adequately only if a *ceteris paribus* requirement is introduced with regard to the side that is not presently under consideration: when evaluating user conduct we must hold the

[37] I will say more about methods of and measures for evaluation in section three of this chapter.

system constant; when evaluating the system we must hold user conduct (somewhat) constant.[38]

Since I aim to evaluate social media *services*, the present work is an instance of supply-side systems-oriented social epistemology.[39] A more precise definition of this type of epistemology is therefore warranted. Supply-side epistemology is a kind of systems-oriented social epistemology that evaluatively assesses the epistemic environment with which individuals are faced in the context of a static social epistemic system. The term "epistemic environment" (henceforth: EE) is commonly used in work on phenomena like echo chambers, fake news, and trust networks, but it is rarely made explicit what the term is supposed to mean. In Michael Baurmann's work, the EE is said to contain institutions and information that affect the outcome of people's individually rational strategies.[40] C. Thi Nguyen is explicitly interested in *hostile* EEs, which he likens to "intentionally placed minefields, but also crumbling ruins, the deep sea, and Mars."[41] Features of hostile EEs attack our (epistemic or cognitive) vulnerabilities. Even Shane Ryan, who calls himself an 'epistemic environmentalist', shies away from a clear definition of the term, although he does observe that the notion of EEs goes hand in hand with the notion that epistemic agents are (socially) situated.[42] Extending this idea, I conceive of the epistemic environment as the constructed[43] environment with which a situated epistemic agent is faced: it is the *social situation* that faces the epistemic agent.

Supply-side epistemology is a strict subset of all work on EEs: whereas all supply-side epistemology is concerned with EEs, not all EEs can be studied by supply-side epistemology. This is because some EEs are dynamic: a trust-network, for example, is defined by the roles that different people occupy in that network. Trust-networks are

[38] That is, if one's evaluative standard is consequentialist. According to Goldman, SYSOR SE usually employs a form of epistemic consequentialism (Goldman, "A Guide to Social Epistemology," 14.); also see section three of this chapter. Under other types of evaluation, a *ceteris paribus* requirement might not be as necessary because it is less important to isolate the consequences of a specific intervention in the system.

[39] Henceforth: supply-side epistemology.

[40] Michael Baurmann, "Fundamentalism and Epistemic Authority," in *Democracy and Fundamentalism*, ed. A. Aarnio, The Tampere Club Series (Tampere: Tampere University Press, 2010), 48.

[41] C. Thi Nguyen, "The Seductions of Clarity," *Royal Institute of Philosophy Supplement* 89 (May 2021): 229, https://doi.org/10.1017/S1358246121000035.

[42] Shane Ryan, "Epistemic Environmentalism:," *Journal of Philosophical Research* 43 (2018): 109, https://doi.org/10.5840/jpr201872121.

[43] "Constructed" is here opposed to "naturally given": the epistemic environment does not consist of trees and rocks, but rather of newspapers, universities, trust networks, and street signs.

therefore studied by dynamic SYSOR SE, even though they are part of the epistemic environment.

## 2. Establishing user conduct

I mentioned in the previous section that supply-side epistemology requires a ceteris paribus requirement on the demand side: the conduct of social media users must be held relatively constant. This raises the question: what sort of behaviour ought I stipulate?

As I mentioned in the introduction, this work is partially inspired by social concerns: there are real worries that the designs of social media services have adverse societal effects. In order for this work to optimally engage this wider societal debate, I will therefore postulate agents that I deem to be as close to actual humans as possible. In making this assumption, I align myself with (at least) two important contemporary philosophical traditions. I just discussed the first: environment-oriented social epistemology. Nguyen, for example, describes his work as "a study in the vulnerabilities of limited, constrained cognitive agents". The second philosophical tradition that makes similar assumptions is non-ideal political philosophy. Contemporary political philosophy is divided over the question to what extent the behaviour of citizens can be idealised.[44] Specifically, this debate revolves around the question whether full compliance can be stipulated: should the theory explicitly deal with non-compliance and illegal behaviour? This can be seen as analogous to the question whether an epistemological theory can assume full compliance with rational rules. Thus, adopting the terminology from political philosophy, we might call the present work an instance of non-ideal social epistemology.

What is a non-ideal epistemic agent (NIEA)? I postulate four characteristics. NIEAs are *finite*; they employ certain *cognitive heuristics*; they are *psychologically situated*; and their societies are characterised by a *division of epistemic labour*.

The finitude of NIEAs concerns their cognitive capacities: there is only so much a single NIEA can remember; there are only so many cognitive processes which a single NIEA is able to run (at a given time and in general); *and* such cognitive processes take time.

This finitude necessitates the use of certain cognitive heuristics: quick and dirty shortcuts to reduce one's cognitive workload. Such heuristics can both be evolutionarily

[44] Ingrid Robeyns, "Ideal Theory in Theory and Practice:," *Social Theory and Practice* 34, no. 3 (2008): 341, https://doi.org/10.5840/soctheorpract200834321.

hard-wired and instances of learned behaviour. We can expect these heuristics to be generally reliable: a heuristic that gets things consistently wrong will only increase one's cognitive workload. Yet they are not consistently reliable, which makes them a vulnerability in hostile epistemic environments.[45]

NIEAs come with an evolutionary inheritance that goes beyond their epistemic capacities: certain facts about human psychology play a role in the (epistemic) behaviour of NIEAs. Thus, to understand the epistemic behaviour of NIEAs, we must take into account what we know about human psychology.

Lastly, the finitude of NIEAs means that life in a complex contemporary society must go hand in hand with a division of epistemic labour. In its simplest and most important form, this means that knowledge itself is divided amongst portions of the population: doctors know medical stuff, engineers know bridge-building stuff. Due to this, our ability to navigate society is largely dependent on our ability to locate the best sources of information on a given topic.[46] More generally, our own doxastic attitudes are dependent on the epistemic standings of others: I know how to deal with my eczema because my doctor told me; she herself read about it in an academic paper; that paper went through a number of checks and edits performed by people specialised in that area; et cetera. Thus the accuracy of my own beliefs is dependent on whether other people in my community do their epistemic parts accurately.[47]

This last point is of course less a fact about NIEAs themselves, and more about the epistemic environments in which they operate. I nevertheless mention it, because it brings to the fore a last important characteristic of NIEAs which I will assume throughout this work: they are not researchers, nor do they usually engage in epistemic activities for non-instrumental reasons. Rather, they approach the world as epistemic agents *because that is how humans must approach the world*. A modern human who wishes to stay alive *must* participate in the division of epistemic labour because she *needs* accurate information about all kinds of things. And this is the crux of the issue: gathering

[45] For a (semi-)recent overview of the state of psychological research in the area of heuristics, including discussion of the reliability of such heuristics, see Gerd Gigerenzer and Wolfgang Gaissmaier, "Heuristic Decision Making," *Annual Review of Psychology* 62, no. 1 (January 10, 2011): 451–82, https://doi.org/10.1146/annurev-psych-120709-145346.

[46] Sanford Goldberg, *To the Best of Our Knowledge: Social Expectations and Epistemic Normativity*, First edition (Oxford, United Kingdom: Oxford University Press, 2018), 242–43.

[47] For an extensive account of the notion of the division of epistemic labour, see Sanford Goldberg, "The Division of Epistemic Labor," *Episteme* 8, no. 1 (February 2011): 112–25, https://doi.org/10.3366/epi.2011.0010.

information is usually merely instrumental to achieving some other goal, so the less time it takes to gather information, the more time remains for that thing which the NIEA really wanted to achieve. From this perspective, a well-ordered epistemic environment is an environment which smoothens this process: it is an environment that allows NIEAs to navigate the division of epistemic labour in order to accurately and easily acquire the information they need.

### 3. Goldman: epistemic evaluation

This brings me to another important question: what measure will I use to evaluate epistemic systems? I just provided a preliminary answer: the extent to which an epistemic environment (or system) allows an NIEA to navigate the division of epistemic labour is the extent to which it is epistemically good. But the question deserves some further consideration.

First of all, what does it mean to navigate the division of epistemic labour? Underlying the phrase is a spatial metaphor: human society is a landscape, and human knowledge is distributed across it. To navigate the division of epistemic labour is to know where to go for which information: whom to ask for help with plumbing, whom for medical advice. Doing this well does not merely depend on finding people who are willing to talk about these matters, but rather on finding people who have actual knowledge. For this is the eventual goal: acquiring the knowledge one needs in order to live one's life.

This shines some light on *why* epistemic environments need to aid in navigation: they need to do so because it will increase general knowledge possession. To accommodate this idea I adopt an epistemic consequentialist framework, as is common in SYSOR SE,[48] broadly along the lines of Goldman's *veritism*. The guiding question of veritistic social epistemology is to what extent a given social structure has "a comparatively favorable impact on knowledge as contrasted with error and ignorance".[49]

In his articulation of veritism, Goldman uses a thin conception of knowledge as true belief. An advantage of such a conception is that it keeps things simple: by merely focusing on the promotion of true belief, complicated questions about justification can be avoided. A disadvantage, on the other hand, is that the simplicity comes with limited explanatory power. For the present project, I find the thin conception of knowledge wanting because

[48] Goldman, "A Guide to Social Epistemology," 14.

[49] Alvin I. Goldman, *Knowledge in a Social World*, Reprint (Oxford: Clarendon Press, 2003), 5.

it lacks a conception of epistemic responsibility. Since the central question of the present work regards the epistemic responsibility of social media services, I need to look beyond Goldman's philosophy in order to strengthen my knowledge conception. Luckily, I do not need to look far: the work of Sanford Goldberg, who stands in the same reliabilist tradition as Goldman, offers a knowledge-conception which both integrates the social dimension at its very core, and incorporates a strong notion of epistemic responsibility.

### 4. Goldberg: knowledge and epistemic responsibility

Goldberg's conception of knowledge is deep and complicated. In what follows, I provide a (very) condensed version of it. Considerations of space and relevance force me to omit many nuances: whenever I am forced to simplify matters extensively, I therefore provide a reference to the relevant parts of Goldberg's account.

Goldberg argues that there are two kinds of considerations that go into assessing whether a belief amounts to knowledge. First, the belief needs to be epistemically proper – I will say more about this soon. On top of epistemic propriety, the belief needs to satisfy a number of non-epistemic conditions in order to be knowledge.[50] Specifically, Goldberg notes two such conditions: the belief needs to be true, and there must be an absence of Gettier conditions.[51]

Goldberg's main interest is in characterising epistemically proper belief. His use of the word "propriety" reflects his interest in the normative dimension of epistemic assessment. Goldberg argues that "S's belief that p is epistemically proper if and only if it enjoys *ultima facie epistemic justification*."[52] The theory of epistemic propriety is hybrid: it evaluates belief on two separate dimensions. First, the belief is evaluated on two core criteria: these are the distinctively epistemic standards of knowledge. Then, the belief is evaluated on the satisfaction of certain general expectations, which are dependent on social context. Goldman argues that such hybrid forms of evaluation are quite common. For example, a similar evaluative structure is used when assessing a candidate for a certain job. There are a number of core criteria which the candidate needs to meet: the candidate needs to have the right knowledge and skills. But there are also expectations

[50] Goldberg, *To the Best of Our Knowledge*, 75.

[51] Goldberg understands such conditions in terms of epistemic luck: see Sanford Goldberg, "Epistemic Entitlement and Luck," *Philosophy and Phenomenological Research* 91, no. 2 (September 2015): 273–302, https://doi.org/10.1111/phpr.12083.

[52] Goldberg, *To the Best of Our Knowledge*, 17.

that come with being an employee in general: regular bathing is appreciated, for example, and a certain degree of sociability. Meeting these general expectations does not count in favour of a candidate: no-one was ever hired because they bathed regularly. But failing to meet these expectations can disqualify an otherwise good candidate. The case of epistemic propriety is an epistemic application of this more general kind of two-dimensional evaluation.[53]

The core criteria of epistemic propriety, also called the distinctly epistemic standards on knowledge, are twofold. The first requirement is that the belief needs to be formed through a process on which the subject is permitted to rely. This permission to rely is dependent on the reliability of the process: if a process is sufficiently reliable, in the sense that its ratio of true-to-false outputs passes a certain threshold, then the subject enjoys a permission to rely on that process.[54] The second requirement is that the belief needs to pass a coherence check with the subject's background beliefs. If both these requirements are met, then the belief is called *prima facie epistemically proper*.

Interestingly, Goldberg argues that these epistemic standards "are what they are *precisely because we are entitled to have certain epistemic expectations of other subjects*".[55] The idea is this: we are epistemic creatures, with an interest in acquiring knowledge. We are also intrinsically and ineliminably epistemically dependent on one another. Based on these facts, we are entitled to expect others to meet certain minimal epistemic standards, and those are exactly the core criteria for epistemic assessment. On the flip side, these legitimate expectations come with a responsibility to live up to them.[56] Thus the normative force of the distinctively epistemic standards on knowledge derives from our epistemic interdependence.[57]

Whereas the core criteria of epistemic propriety derive from the fact that we are epistemically interdependent subjects, the general expectations derive from the fact that we are all also embedded in specific – and to some degree contingent – epistemic communities. Our social-epistemic positions, the roles we play in specific epistemic

[53] Goldberg, 48–71.

[54] In fact, Goldberg's account is slightly more complicated than this: it distinguishes between processes for which we enjoy a *default* permission to rely, and processes for which we can *earn* permission to rely. To keep the discussion focused, I will disregard this nuance. For more information, see Goldberg, *To the Best of Our Knowledge,* 75-112.

[55] Goldberg, *To the Best of Our Knowledge*, 147.

[56] Goldberg, 183.

[57] Goldberg, 142–43.

practices, come with certain epistemic expectations. Doctors are expected to be knowledgeable about the latest treatments; journalists are expected to meet certain investigative standards. Since these too are legitimate epistemic expectations, they translate into an epistemic responsibility to meet them. It is only when, on top of satisfying the core criteria, the epistemic subject also satisfies the general expectations, that a belief can be called "*ultima facie* epistemically proper", or simply "epistemically proper".[58]

We thus see that it is one's epistemic responsibility to ensure that one, in the formation of one's beliefs, meets the legitimate epistemic expectations that others have of one. These legitimate expectations are those formulated in the core criteria – reliability and coherence – and those that derive from legitimate[59] social practices in which one partakes. When one meets these expectations and therefore satisfies one's epistemic responsibilities, one's beliefs meet the standards set by epistemic propriety. Such are the epistemic responsibilities one has with regards to belief-formation. But these are not the only epistemic responsibilities there are: a second class of epistemic responsibilities derives from the nature of assertion.

## 5. Assertoric responsibility

Goldberg argues that assertion can be defined as the (only) speech act with a constitutive rule.[60] The constitutive rule of assertion is epistemic in nature: one may only assert that p, if one has a specific epistemic standing with regards to p. From this rule derives a legitimate expectation: we can expect others to adhere to the rule. And just like before, from this legitimate expectation arises the responsibility to meet it.[61]

Although Goldberg does not explicitly discuss it, we can see that the legitimacy of the expectations that are called forth by the rule similarly derive from basic facts about our situation as epistemic subjects. In the previous section, we saw that our legitimate expectations with regard to belief-formation derive from two facts: we are epistemic

[58] Goldberg, 148–49. This, too, is slightly more complicated than represented here. For Goldberg, it matters for epistemic propriety whether living up to the general expectations, in cases when one has failed to do so, would have rendered the belief *prima facie* epistemically improper. Once again, I will disregard this nuance at present: for a more detailed account, see Goldberg, 186-227.

[59] Of course, Goldberg stipulates conditions that a practice needs to meet in order to be legitimate. See Goldberg, 169-175.

[60] Sanford Goldberg, *Assertion: On the Philosophical Significance of Assertoric Speech* (New York, NY: Oxford University Press, 2015), 11.

[61] Goldberg, 73–75.

subjects with an interest in finding truth, and we are ineliminably epistemically interdependent. The expectations with regard to belief-formation which were formulated in the previous section can by themselves not entirely ensure that we are lifted out of ignorance: even if everyone forms beliefs reliably, we still require a reliable means to communicate these beliefs. Assertion, a speech act defined by a constitutive, distinctly epistemic rule, helps us meet this last requirement.[62]

While the need for assertion, and therefore the legitimacy of the expectations and responsibilities that come with it, arises from our general situation as epistemic subjects, Goldberg argues that the precise content of the rule depends on our specific epistemic situation: the rule is context-dependent.[63] Thus the normativity of assertion echoes the structure of the normativity of belief formation. Specifically, the epistemic standing which one must have with regard to p in order to assert it is dependent on the mutual belief of the speaker and the audience regarding the audience's interests and informational needs.[64] This is less complicated than it sounds: if the audience requires information in order to address a practical task, then the speaker must only assert propositions for which he has the kind of evidence that would make it rational to base one's course of action on it; if the audience requires absolute certainty, then the bar for assertion is raised accordingly.

Although Goldberg only discusses cases of interactions between individuals, it is clear that assertoric expectations and responsibilities can be institutionalised in complex societies in quite the same way as belief-related expectations and responsibilities can. Think, for example, of the opinion pages in a newspaper: a different epistemic standing is expected of writers with regard to the assertions they make in these pages, than is expected of writers in other parts of a newspaper. By dividing newspapers into opinion and news pages, the question is settled which assertoric norm should be upheld, even though speaker and audience are in no direct contact with each other. Thus assertoric

[62] This mirrors Bernard Williams' state-of-nature-account of the two prime epistemic virtues. We can understand the responsibilities with regard to belief formation as roughly corresponding with the virtue of Accuracy, while assertoric responsibilities roughly correspond with the virtue of Sincerity. See Bernard Williams, *Truth & Truthfulness: An Essay in Genealogy* (Princeton, N.J: Princeton University Press, 2002), 41–62.

[63] Goldberg, *Assertion*, 225.

[64] Goldberg also includes 'the prospects for high-quality information in the domain in question' as part of the relevant mutual belief. He introduces this to address a specific issue in the epistemology of disagreement; for the sake of relevance, I omit it here. See Goldberg, 257.

expectations and responsibilities are another important part of the division of epistemic labour in complex societies.

### 6. Context clues

The picture so far is this. In contemporary, complex societies, populated by non-ideal epistemic agents, knowledge is distributed across the social sphere: I have referred to this as the “division of epistemic labour”. In order to make navigation of this division of epistemic labour possible – in order to make it possible for people to get the information they need – the social sphere is partitioned into an assortment of social practices. Each of these practices is situated in a certain epistemic domain (e.g. medicine, plumbing, engineering, et cetera), and every practice comes with a set of epistemic responsibilities regarding belief formation and standards for the kind of epistemic standing that is required for assertion given a certain context. Thus if one knows the social practice in which one’s interlocutor is situated, one will know the kind of knowledge one’s interlocutor can be expected to have, the quality of their belief-formation processes, and the conditions under which they will make assertions. This knowledge will help one meet the demands of epistemic propriety when it comes to one’s own belief-formation processes.

This raises the question: how *does* one know in which social practice one’s interlocutor is situated? In many situations, one simply knows: it is part of one’s general knowledge who one’s doctor is, or what the country’s most important newspapers are. Even in new situations – for example, after a move to a new city – we navigate the division of epistemic labour with relative ease by observing the epistemic environment: the person behind the door with the word “doctor” on it can be expected to be situated in the social practice called “medicine”; the thing that fell on the doormat this morning looks like a newspaper, so it probably fits in the social practice called “journalism”. Such heuristics can only be stretched so far: the further we are from the environment that is familiar to us, the less effective we will be at navigating the division of epistemic labour, for our heuristics are attuned to the social situation in which they took shape. The heuristic that recognises ‘having a degree from a medical university’ as a marker for ‘being a trustworthy source of information regarding my skin condition’ only applies in a context where medical universities are qualitatively good institutions: it is not universally reliable.

This kind of context-sensitivity does not only apply to heuristics that allow us to navigate the division of epistemic labour: the same goes for many other heuristics and cognitive processes. Goldman defines reliability in terms of performance in normal conditions. This means that if the conditions change, then the reliability of our cognitive processes might also change. An example of such a context-sensitive cognitive process is the fluency heuristic, a psychological phenomenon that causes people to assess information that is easier to process as more likely to be true.[65] Reber and Unkelbach argue that the reliability of this heuristic is context-sensitive: it is reliable only in situations in which the majority of statements to which one is exposed is true.[66]

### 7. Media and epistemic propriety

In my discussion of the division of epistemic labour in contemporary society, I have so far disregarded one important aspect: media. It is undeniable that media play an important role in (our capacity to navigate) the division of epistemic labour in contemporary society. Important information – for example regarding emergency measures in the face of a pandemic – reaches the population through the channels of the mass media. Yet it is hard to see how any of the aforementioned responsibilities can apply to media organisations.[67] The responsibilities with regard to belief formation do not apply to media, because media are not epistemic subjects: they are not the type of entity that is capable of doxastic states. Assertoric responsibilities, on the other hand, are not applicable because media *themselves* generally do not make assertions: rather, they carry assertions which were made by others to a wider audience.

When we reconsider the preceding accounts of epistemic responsibility, we realise that epistemic responsibilities arise from epistemic abilities. Specifically, responsibilities

[65] A more in-depth account of the fluency heuristic is provided in chapter six.

[66] Rolf Reber and Christian Unkelbach, "The Epistemic Status of Processing Fluency as Source for Judgments of Truth," *Review of Philosophy and Psychology* 1, no. 4 (December 2010): 563–81, https://doi.org/10.1007/s13164-010-0039-7.

[67] In what follows, I will speak as if media themselves can have responsibilities. This is not meant in any literal sense: I do not believe that mere technology can bear responsibility (epistemic or otherwise). The responsibility accrues, instead, to those responsible for the technologies. This passes the buck, of course, because it raises the question who or what is responsible for these technologies. Is it the people who create the technology? Those who profit from it? Or might it perhaps be the corporate entities that own the technology? This is a complex set of questions, the answering of which would take me far beyond the scope of the present work. For an interesting take on the topic of corporate moral responsibility based on broadly epistemic considerations, see Philip Pettit, "The Conversable, Responsible Corporation," in *The Moral Responsibility of Firms*, ed. Eric W. Orts and N. Craig Smith (Oxford University Press, 2017), 15–35, https://doi.org/10.1093/oso/9780198738534.003.0002.

arise from the power each of us wields over the shared pool of human knowledge. The responsibilities regarding belief-formation are necessary due to the fact that others can gain access to our beliefs through testimony; the responsibilities regarding assertion are due to the fact that humans can assert – and thereby affect other people's belief formation processes. But media can also wield considerable epistemic power, although it is of a different kind. Rather than wielding power over belief-formation or assertion, media wield curatorial power: the power to decide who sees which content at what time. Curatorial power constitutes the ability to affect the context within which new evidence is introduced. We just saw that context is extremely important for our ability to interpret evidence reliably, so the ability to affect context equals an ability to affect the belief-formation processes of others.

Consider media-broadcasts of assertoric speech. Remember that the constitutive rule of assertion is context-dependent: whether I may assert *p*, given that I have a certain epistemic standing towards *p*, depends on the context in which I wish to assert it. Specifically, it depends on my and my audience's mutual beliefs about my audience's informational needs. This means that subject S might responsibly assert *p* toward audience $H_1$ given that S's standing toward *p* is *x*. Now let there be another audience, $H_2$, whose informational needs are different to such an extent that S would need epistemic standing *y* toward *p* in order to assert it. S does not have standing *y*, but *x* toward *p*. Thus S can legitimately assert *p* toward $H_1$. Unbeknownst to S, however, her assertion is recorded, and is broadcast by media to $H_2$. In this case, it would be epistemically improper for $H_2$ to accept *p* on the basis of S's say-so, because S lacked the required epistemic standing vis-à-vis *p* in order to assert it toward $H_2$. But the broadcasting media can make it so that there is no way for $H_2$ to know this, making it likely that the assertion will be wrongfully accepted.

A similar problem holds for broadcasts that do not distribute assertions, but records of other events. Whether it is video material of war zones or recordings of the sounds of the forest: such records are lifted out of their 'natural' environment, thereby making it hard to assess their significance. The changed context might affect the reliability of inductive reasoning, for instance, if the selection process that preceded broadcast was biased in some way. Thus the change of context that media can introduce – by omission or otherwise – can affect our ability to meet the demands of epistemic propriety.

## 8. Media and responsibility

If we wish to avoid wholesale epistemic impropriety due to the introduction of media into society, then the relationship between epistemic responsibility and curatorial power must be explicated. This explication has two purposes: first, it serves to explain the social practices that have been erected around traditional media like newspapers and television; second, grasping the rationale underlying these social practices will help us understand what the relationship between social media and epistemic responsibility might look like.

Media allow audiences to perceive events from a distance. Importantly, this perception at a distance is mediated by some degree of human involvement: a medium[68] is not merely a dead tool, like a set of binoculars or a pair of glasses. Perception by means of television, for example, is mediated by an extensive process of editing: the same goes for radio and newspapers. There is no requirement as to how active this human involvement must be. For example: a tv-station decides to aim a camera at a city square, and live-stream whatever that camera records for the next twenty-four hours. Even if the crew does not intervene at all during those twenty-four hours, just allowing the technology to do the work on its own, the tv-station remains a medium: after all, it was the decision of those working at the tv-station to broadcast like this. The difference with the case of binoculars is that a third party retains curatorial power over the process of perception, even though that power is not used very actively in this particular case.

In some cases, this curatorial power can be called epistemically significant. Television is an example: a camera crew can, by focusing the camera on this rather than that, lead the audience astray about what is going on. The same goes for radio: by selecting what to play, the DJ holds sway over the belief formation processes of her audience. The camera crew and the radio DJ wield epistemic power.

Other media are structured in such a way that it is clear how curatorial power is wielded, which minimises the epistemic effects of these media. Think of telephone companies: although their activities also facilitate their customers' perception at a distance, the way the phone network is organised means that it remains largely in the hands of the customers what they perceive. The same holds for the postal services: the activities of the postman do not influence how the letters he delivers are interpreted. I

[68] For clarity's sake I will depart from normal usage, using "medium" as a singular of "media".

will call media like the telephone and postal service *transparent*, and media like television and radio *opaque*.

The term “transparency” might, at first, call forth the idea of an unobstructed view. The medium, on this interpretation, is a barrier between the audience and the event, and if the barrier is transparent, then the audience can perceive the event as if the barrier was not there. But this is a misleading way to understand the transparency condition, for even transparent media are selective in what they transmit: the telephone does not give an “unobstructed view” of the event it transmits, for visual information is lacking, and even only a small portion of all auditory information is transmitted. Rather than intending “transparency” in the sense of an unobstructed view, I intend to use the term in the proverbial sense as a way to describe the activities of others: a person or an organisation is transparent in this sense if it is open to the public about its inner workings. If a medium is open in this way, then it is clear from its structural logic *in which way* it might misrepresent events, thereby salvaging the audience’s ability to properly interpret what it perceives through it.

This is the epistemic significance of transparency. The fact that the audience’s ability to properly interpret what it perceives through transparent medium M is salvaged, means that perception-through-M is a reliable process – as long as the audience takes into account the (predictable) ways in which perception is obscured by the medium. Transparent media are epistemically *good*, in the sense defined in section three: they promote knowledge possession. A non-transparent (opaque) medium M’, on the other hand, is epistemically problematic because there is no way to tell in which way perception through M’ might be misleading, and therefore there is no way to tell in advance whether perception-through-M’ will be reliable. Thus if we still wish to engage epistemically with what we perceive through M’, then a new set of epistemic standards is necessary to ensure that this process happens within the limits of epistemic propriety. Luckily, Goldberg provides for a way to institute such standards: by establishing a social practice.

This, I think, goes some way toward explaining the proliferation of social practices around traditional media. The best example is journalism, an epistemic practice that regulates a subset of all activities on a variety of media. This practice puts arduous epistemic demands on those who call themselves journalists, both with regard to their research activities and with regard to what they may assert. Media organisations which are situated in this social practice are expected to ensure that the assertions they broadcast are made by people who meet the epistemic demands implied by this practice.[69]

At a higher degree of generality, slander and libel laws put demands on all media. Working within this frame are numerous media organisations that further clarify which expectations their audiences may legitimately have of them by profiling themselves in this or that way: a socialist magazine might be expected to skew stories in one direction, a Christian tv-channel in another. The epistemic responsibilities of audiences walk in step with these legitimate expectations: it is more reasonable to form beliefs on the basis of a single purely journalistic article than it is to form beliefs on the basis of a single article from a socialist magazine, since it is part of the social practice called "journalism" to attempt to reach a balanced conclusion after hearing all the sides of the story, while a magazine with an explicit political orientation can be expected to emphasise one side over the other.

Some media, like gossip magazines, situate themselves in a social practice that is even less epistemically demanding. In such cases, the brunt of the epistemic responsibility falls on the shoulders of the audience: it is distinctly epistemically improper to draw conclusions merely based on the reporting of gossip magazines. Thus refraining to put substantial epistemic demands on the social practice that surrounds one's activities as a medium comes at a high cost: it renders the medium practically incapable of playing a role in processes of belief-formation.

### 9. Social media and epistemic responsibility

With a clear account of the relationship between epistemic responsibility and media in hand, we can now return to the central question of this work: what are the epistemic responsibilities of social media services? Much hangs on whether social media services are transparent in the way just described. This question remains thorny, however, for while it is clear what the *effects* are of a transparent medium, there is still a variety of ways in which transparency may be achieved. Thus evaluating the transparency of social media is not simply a matter of ticking the boxes: in order to see whether the cognitive processes in which the broadcasts of social media play a role remain reliable, a proper evaluation of the reliability of these processes is required. This is the topic of chapters

[69] For a more elaborate discussion of journalism, see chapter four.

four, five, and six. Nevertheless, I am already in a position to make some general remarks about why social media services might or might not be expected to be transparent given the way they are designed.

Looking at the historical development of social media services as outlined in chapter one, there is a clear reason to expect that the transparency condition may have been fulfilled by the earliest social network sites. These sites were not focused on the distribution of content as much as they were focused on mirroring existing (off-line) social networks in an online environment in order to ease communication. In such a network, the transparency condition might be expected to be fulfilled because the network only facilitates communication in a context that is already familiar. The initial focus on direct messaging made these sites something like an online analogue of the postal system, which I argued is itself a transparent medium. I believe this to be the intuition which underlies the adage that social media services are "platforms", and therefore neutral. Of course, more elaborate research would be required to establish whether this intuition is correct.

Such research would be of little help for those who wish to understand the responsibilities of contemporary social media, however, for – as I showed in chapter one – social media services have come a long way from the social network sites of old. The biggest wild card that is introduced are recommender systems. As I argued in the previous chapters, the design of these systems is an expression of the curatorial power which these services wield. Is this power of the right kind to be characterised as epistemically significant – and therefore to cause contemporary social media services to be characterised as opaque? This is the question which I aim to answer in the four chapters which constitute part two of this work.

Here is one reason why even contemporary social media services might be transparent. In chapter one, I mentioned that recommender systems are commonly understood as systems that aim to recommend to users content that they will like. If we take this conception seriously, then recommender systems function by tracking the preferences of users. But if *that* is correct, then these systems merely smoothen a process that would have occurred anyway: yes, the system presents user H with content *c*, but since the system merely tracks users' preferences, it might be reasonably expected that H would have found *c* anyway. The only difference, then, is that the system has made it easier for H to find c: rather than having to find a magazine that fits his interests, the system proactively provides H with the same content that the magazine would have had. If this is how the system functions, then it might indeed be transparent, for it does not change much apart from making life a little easier for H.

There is, I think, much wrong with this argument-from-preference-tracking. But rather than addressing it here, and head-on, I will take it as a guiding idea for the second part of this work.

# Part two: Epistemic Effects

## The lay of the land

Social media are online services offering content-functionality, combined with some form of automated and personalised curation. The way curation is performed tells us how social media services wield their curatorial power, which is the power to decide who sees which content at what time. The main criterion on the basis of which curation is performed by contemporary social media is attention maximisation: social media recommender systems aim to increase the total time spent on the service by users.

The epistemic responsibilities of media depend on a number of factors. Media can play an important role in (our ability to navigate) the division of epistemic labour: well-ordered media make it easier for citizens to acquire accurate information. Such well-ordered media are epistemically *good*, in the sense that they increase general knowledge possession. Some media are well-ordered by nature, because they are transparent: it is clear to audiences how these media distort the information they pass on, which allows audiences to correct for the distortions in their processes of belief formation. Other media are opaque: because these media do not wear on their sleeves what kind of distortions they introduce, it is impossible to tell in advance the reliability of belief-formation processes which are based on their broadcasts. Opaque media are epistemically problematic, unless a social practice is erected around them that puts legitimate epistemic expectations on media and those who interact with them.

In order to establish the epistemic responsibilities of social media, we first need to answer the question whether social media services are transparent in the way just described. If they are, then no particular responsibilities attach to them, because they are already epistemically good. If social media services are opaque, then social practices must be erected in order to allocate epistemic responsibility. This leaves open the question *how* this responsibility must be allocated. On one extreme, all responsibility falls on the shoulders of the audience. In such a case, the epistemic responsibilities of the audience will likely mirror those of the audiences of gossip magazines. On the other extreme, the social media service itself must shoulder strong epistemic responsibilities in order to protect the reliability of its audience's cognitive processes.

In the remainder of this thesis, I start filling in the finer details of my account of the epistemic responsibilities of social media by tracing the epistemic effects of the way social media services wield their curatorial power. I argue that attention maximising recommender systems can be expected to have certain epistemic effects, and that therefore the curatorial power of social media services can be characterised as epistemically significant. Indeed, the choice for attention maximisation can be epistemically detrimental: it can negatively affect the reliability of the audience's belief formation in unpredictable ways. For these reasons, I argue, social media services are opaque.

### The road ahead

The second part of this work, comprising of chapters four, five, six, and seven, provides an argument by means of examples. By studying three different cases I show that social media services are not transparent media, and that some epistemic responsibilities must fall upon the shoulders of the service itself. The three cases correspond to the three attention-maximising strategies I identified in chapter two: the hook, the line, and passive attention drawing.

For each of these strategies, the functioning of the strategy is contingent on the behaviour of a user being predictable in some way: a hook can only function as a hook if it is knowable in advance that a certain user will be impulsively drawn to certain content in a certain situation, the same goes for lines and passive strategies. But this leaves unaddressed which mechanisms underly this predictability. One popular account of such a mechanism, which is often heard in the context of recommender systems, is that the recommender system learns to track the user's preferences: once the system finds out that Mark is a cat lover, it can predict with relative certainty that a cat video will draw Mark's attention. I argued in chapter three that a recommender system's latching onto a user's preferences in this way might provide a foundation for an argument to the effect that social media are transparent.

In **chapter four**, *The Hook: Deceptive News*, I argue that the argument-from-preference-tracking is wrong: in many cases, recommender systems do not track users' preferences at all. By analysing the functioning of the news market on social media I show that the account of preference tracking rests on a confusion about the nature of preferences. While a person's behaviour might seem to manifest a preference for fake news, this preference only manifests due to the unfortunate epistemic environment which the social media service offers. This causes recommender systems to sometimes recommend content that is diametrically opposed to a user's preferences. In the case of the news market, this means that the system might recommend fake over real news.

In **chapter five**, *The Line: Rabbit Holes*, I provide an account of one way in which a recommender system might aim to sustain attention once it has been grabbed. This account illustrates the sheer complexity of users' behaviour on social media services: I argue that such behaviour might be predictable due to an interplay between epistemic capacities, certain psychological facts about users, and the epistemic virtue of curiosity. A recommender system can use this predictability to its advantage, using content of a certain type to stimulate prolonged interaction with the platform.

On top of showing the incorrectness of the argument-from-preference-tracking, chapters four and five also provide examples of ways in which social media services might be opaque: the interventions of the recommender system are epistemically detrimental. In **chapter six**, *Passive Attention Drawing: Belief-Driven Strategies*, I argue that the manipulation of our epistemic activities in which recommender systems engage is much less innocuous than the argument-from-preference-tracking portrays it. Epistemic activities tend to be aimed at forming new beliefs, and manipulation of these activities is therefore likely to entail a manipulation of our beliefs. This means that manipulation of our beliefs themselves comes into recommender systems' purview as an additional strategy for attention maximisation.

In **chapter seven**, *The Epistemic Responsibilities of Social Media Services*, I draw my conclusions from the preceding chapters. First, I argue that social media services are opaque. Second, I argue that mere responsibilities on the side of a social media service's audience will not suffice if we wish to salvage the societal role that social media services presently fulfil. I conclude by offering some thoughts on how the epistemic responsibilities of social media services can be further concretised.

# Chapter Four

## The Hook: Deceptive News

Campaign cycles in the United States are notoriously ferociously fought. The 2016 cycle, however, saw a new phenomenon altogether. In the months leading up to the election, strange stories started entering public debate: presidential candidate Hillary Clinton had sold weapons to ISIS, an FBI agent suspected of leaking Clinton's emails had been found murdered, and pope Francis had endorsed presidential candidate Donald Trump. These stories had two things in common: they had all garnered enormous attention on Facebook, and they were all false. Indeed, in the three months leading up to the November election, the twenty best performing fake news election stories generated more Facebook engagement than the twenty best performing real news election stories.[70] What was going on here?

Since 2016, much has been written about the fake news phenomenon. In the present chapter I build on that work, but take a different perspective. Using the news market on social media as a test case for the hypothesis that recommender systems track users' preferences, I show that in a deficient epistemic environment, a person's actual preferences can diverge considerably from the preferences which seem to manifest from that person's behaviour. Because recommender systems can only deduce a person's preferences from the way they behave, allowing such a system to further shape such an already deficient epistemic environment will likely cause that epistemic environment to deteriorate even further. In the case of the news market, recommender systems' failure to track user preferences reduces users' ability to navigate the division of epistemic labour.

I start the chapter with a discussion of the relationship between recommender systems and preferences. With those insights in hand, I move on to an analysis of the market for news, where truth-preference leads to a paradox which can only be solved by means of certain heuristics. At that point, I will be able to provide an account of what can be expected when the market for news moves to social media – and how this unfortunate marriage leads to the rise of fake and otherwise deceptive news.

[70] Craig Silverman, "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook," BuzzFeed News, November 16, 2016, https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

## 1. Preferences in context

In August 2012 Eric Meyerson, head of content creator communications at YouTube, released a blog post called “Why We Focus on Watch Time”[71] that provides a rare insight into the workings of and ideas behind social media recommender systems. In the post, Meyerson explains that since the average household watches several hours of tv per day, YouTube still has “a lot of growing to do”. To facilitate this growth, Meyerson explains, YouTube has “updated what we call video discovery features, meaning how our viewers find videos to watch via search and suggested videos.” Although Meyerson does not use the exact word, it is clear that these discovery features are driven by algorithms of a kind which is by now well known to us: recommender systems.

Specifically, what YouTube changed in 2012 was the ODS of its recommender system. Before the change, the recommender system optimised for “views”: merely opening the URL of a video was considered a successful engagement, regardless of whether the video was actually watched. After the change, the recommender system had been adjusted to focus on videos “that increase the amount of time that the viewer will spend watching videos on YouTube, not only on the next view, but also successive views thereafter”. Clearly, this change of strategy fits well with YouTube’s professed growth goals. Yet Meyerson says that this is not the only reason for the change. According to him, “If viewers are watching more YouTube, it signals to us that they’re happier with the content they’ve found.” In other words: the user is believed to show a preference for a certain kind of content by engaging with it for an extended period of time.

I will call the preferences which the YouTube algorithm is said to track a person’s manifest preferences: they are the preferences that manifest in a person’s behaviour on YouTube’s service. Opposed to this notion of manifest preferences are what I will call a person’s actual preferences: those preferences which a person would avow to when asked. This distinction between two kinds of preferences should not be taken as a philosophically interesting statement about the nature of preferences, but rather as a simple consequence of the idea that we can sometimes be misguided in our beliefs about what course of action is warranted in light of our desires. It is in these moments of confusion that our manifest preferences diverge from our actual preferences: we might

[71] Meyerson, “YouTube Now.”

manifest a preference for A while under the impression that we are manifesting a preference for B.

Another reason why our manifest preferences might diverge from our actual preferences is because we harbour un- or subconscious preferences that, although we cannot avow them, might nevertheless manifest in our behaviours.[72] In the present work, however, I will ignore this possibility, assuming instead that all our preferences are avowable. This, too, should not be taken as a philosophically interesting statement in the philosophy of mind, but as a tool to simplify the discussion. I take it to be uncontroversial that there are situations in which a divergence between manifest and actual preferences is due to something else than the presence of un- or subconscious preferences, and my interest is in these cases at present. The assumption that all preferences are avowable is thus not an impossibility claim in the philosophy of mind, but rather a way to disregard certain cases which are not relevant to the present discussion.

The manifest preferences that a recommender system tracks are deduced from a subset of a person's behaviours, specifically that person's behaviours on the service. This notion of preferences is loosely based on the role preferences play in some economic theories, where a decision to buy a certain good is an expression of a preference for that good. The preference-based interpretation of YouTube's algorithm is analogous: the decision to watch a video to the end is interpreted as an expression of a preference for that video.

The notion of manifest preferences is severely limited. It is based on an extremely simple conceptualisation of human psychology, according to which all behaviour is a direct expression of a preference. But this conceptualisation is not even sufficient to describe all normal behaviours in everyday situations. Think of a supermarket customer who intends to buy white rice, but accidentally grabs a bag of brown rice because the price tags under the products have been switched: does she manifest a preference for brown rice? Alternatively, the customer might only buy brown rice because she does not know the brand of white rice that is available and therefore fears she might not like it – even though, if she did try it, she would learn that she prefers it. Or lastly, she might buy brown rice because she believes that it is healthier – even though, in her specific situation, it would actually be healthier to eat white rice.

[72] This line of thought is developed extensively in Richard Moran, *Authority and Estrangement: An Essay on Self-Knowledge* (Princeton, New Jersey: Princeton University Press, 2001).

In each of the sketched situations, the customer manifests a preference for brown rice while her actual preference is – or ought to be – white rice. The mismatch between manifest and actual preference in each case is due to gaps in the available information, or faults in the epistemic environment with which the customer is faced: the signs are switched, the customer lacks a method of establishing how her preferences translate into the real world, or she has simply been given the wrong information about healthy nutrition. Thus a deficient epistemic environment can cause a person to manifest preferences which are opposed to her actual preferences.

## 2. Preferences on the news market

With an account of preferences in hand, we can turn to the market for news. To understand how the epistemic environment of social media services influences the functioning of the news market, we must first understand how that market functions outside of that environment. There are three elements to this: we need an account of what news is, we need an account of what customers' preferences are on the news market, and we need an account of how customers manage to satisfy those preferences.

Kay Mathiesen defines news as the product of journalistic practice. Journalistic practice, in turn, is defined by Mathiesen as "the activity of gathering, assessing, creating, and presenting [current] information, where 'getting it right' is the foundation upon which everything else is built."[73] This definition is especially useful in the present context because it underlines the epistemic commitments which news inherently has: "Journalism's first obligation is to the truth."[74] We thus see that journalism is a social epistemic practice in Goldberg's sense:[75] it is a practice from which derive legitimate epistemic expectations. These expectations translate into epistemic responsibilities on the side of journalists: in order to fulfil their social-epistemic role well, journalists must meet the epistemic expectations which others have of them.

[73] Kay Mathiesen, "Fake News and the Limits of Freedom of Speech," in *Media Ethics, Free Speech, and the Requirements of Democracy*, ed. Carl Fox and Joe Saunders, Routledge Research in Applied Ethics 13 (New York: Routledge, 2019), 167. Mathiesen paraphrases American Press Institute, 'What Is Journalism?'. www.americanpressinstitute.org/journalism-essentials/what-is-journalism/. Some quotation marks where omitted for legibility.

[74] "The Elements of Journalism," *American Press Institute* (blog), accessed April 1, 2021, https://www.americanpressinstitute.org/journalism-essentials/what-is-journalism/elements-journalism/.

[75] See chapter three for a treatment of Goldberg's account of social epistemic practices.

Given this definition of news, it is not a big leap to say that consumers on the news market have a preference for true over false reports: if journalists openly declare that their first obligation is to the truth, and if consumers explicitly conceive of their behaviours on the news market as behaviours on the *news* market, then consumers must believe themselves to be consuming *true* accounts – just like a consumer on the bicycle market conceives of himself as consuming bikes rather than cars. The expectations around news flow naturally from the epistemic expectations around journalism: news must be true, or at least attempt to get at the truth. This commitment is taken seriously: in the past, the publication of patent falsehoods by news outlets has led to large scandals, often with severe consequences for the journalists and news outlets responsible.[76]

This last point might lead some to object: is it really the case, in what has been called the era of "post-truth", that news consumers care about the truth of news? Is, for instance, Fox News in the United States not infamous for spreading falsehoods, without reaping any repercussions for it?[77] And does the fact that Fox remains popular then not show that news consumers do not have a preference for truth at all? Now, I do not object to the premises of this argument: Fox has indeed been shown to spread misinformation on multiple occasions, and Fox News is indeed still popular. I do not even disagree with the conclusion: the behaviour of Fox consumers does indicate a preference for falsities over truth. Yet I maintain that news consumers have a preference for truth. This road is open to me because I detect a subtle shift in the way we use the word "preferences". When I speak of the preferences of news consumers, I intend to speak of their actual preferences: I mean to say that if we were to ask a news consumer, she would be likely to profess a preference for true news over false news. When we speak of the preferences of Fox watchers, on the other hand, our conceptualisation of preferences slides from the actual kind to the manifest kind: we observe behaviour from an outside perspective, and immediately explain that behaviour in terms of preferences. But it is *our* observation, supplemented by *us* from the outside perspective, that Fox's reports are substantially false. It would be absurd to claim that these people's *actual* preferences were for false news: who would profess that they like to watch a show because it consistently lies to

[76] Matthew Gentzkow and Jesse M. Shapiro, "Competition and Truth in the Market for News," *Journal of Economic Perspectives* 22, no. 2 (2008): 146.

[77] See by way of example Eliza Relman, "Right-Wing Media Has Pushed 3 Completely False Narratives in Less than a Week," Business Insider, April 27, 2021, https://www.businessinsider.com/right-wing-media-fox-news-3-debunked-stories-in-week-2021-4.

them? If watching Fox News could be seen as an expression of an actual preference for false news, then it would be wise for Fox to advertise that all its news is patently false: that is what consumers want, right? But Fox News does not do this. This shows that the question should not be how some consumers can have developed an appetite for false reports, but how those consumers can be so mistaken about the truth.

Kathleen Hall Jamieson and Joseph N. Cappella go some way toward providing an answer to this question. They argue that Fox News manages to convey a narrative to its viewers that frames all those who oppose Fox News and its deeply conservative outlook as untrustworthy liberals. "When these partisans attend to nonconservative media or confront partisans of opposed political beliefs, this buffer insulates them from counterpersuasion."[78] Indeed, because the narrative predicts that liberals will aggressively lie to undermine conservatism, fact-checking of Fox News reporting and counterevidence offered by other news outlets corroborates rather than undermines the narrative espoused by Fox news.[79] So instead of somehow finding a niche market of news consumers who have a preference for false reports, Fox News works tirelessly to construct an epistemic environment in which its viewers can feel justified in believing the views the network espouses. The success of Fox News is not explained by reference to consumers' preference for false reports: it is explained by the fact that Fox News manages to subvert consumers' intuitions for distinguishing the true from the false.

The Fox News case offers a first glimpse at a deeper paradox underlying the idea of a market for news. A prerequisite for any market to function efficiently is that the market works under conditions of perfect information: buyers and sellers need to have all relevant information to make market-related decisions.[80] On the bicycle market, for example, this means that consumers need to be able to establish with relative ease the quality of the bikes for sale, because the quality of a bike is likely to have an influence on the consumer's willingness to pay for that bike. Analogously, then, for a news market to function, consumers need to be able to establish with relative ease the truth of the reports they consume, for we just saw that the truth-value of a report has a great influence on the consumer's preference for that report. Importantly, this establishing of the truth of a

[78] Kathleen Hall Jamieson and Joseph N Cappella, *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment* (Oxford; New York: Oxford University Press, 2010), 237.
[79] Jamieson and Cappella, 238.
[80] Alvin I. Goldman and James C. Cox, "Speech, Truth, and the Free Market for Ideas," *Legal Theory* 2, no. 1 (1996): 19.

report must occur *before the report has been consumed*, just like a would-be bicycle customer must be able to establish the quality of a bike before she buys it. But upon reflection, this requirement is paradoxical. In order to know the truth of a proposition, it is usually necessary to know the proposition. But if a consumer already knows the propositions contained in a news report, then there is no longer a need to consume that report. The requirement comes down to a requirement that a consumer know what a newspaper says before he buys it. Failure to meet this requirement does not merely mean that the market will be inefficient: it means that customers are not able to identify accurate news reports, which undermines the epistemic rationale underlying the practice of journalism. It seems impossible, then, to have a functioning market for news, because the preconditions for the efficiency of that market make the existence of the market superfluous, while failure to fulfil these preconditions renders the market unable to deliver on its most important epistemic promises.[81] But of course, we know that it *is* possible for markets for news to exist, for we have seen them with our own eyes.

So how is the paradox averted? The consumer requires a proxy for truth: some sort of heuristic to tell the consumer that this report is probably true, while that one is not. In the traditional news market, an important such heuristic is reputation. In such a market, there is often a limited number of well-known news outlets competing with each other: there are some newspapers, some tv-stations, and some radio stations. Because the number of outlets is limited, and because such outlets tend to exist for an extended period, over time these outlets will build a reputation. Lies and failures to report the truth will be remembered, as will instances of exceptional reporting. This is possible because news outlets tend to make what Alvin Goldman has called exoteric statements: although the people publishing the statements are specialists (journalists), the statements they make are often easily checkable by laypeople – perhaps not at the very moment of reading, but certainly some weeks, months or years down the line. If a news outlet has a track record of publishing stories that later turned out to be true, the reader "can infer that this unusual knower must possess some special manner of knowing (…) that is not available to them":[82] good journalist practices, for example.

[81] This argument is inspired by Goldman and Cox, 19–20.

[82] Alvin I. Goldman, "Experts: Which Ones Should You Trust?," *Philosophy and Phenomenological Research* 63, no. 1 (2001): 107.

Of course, I do not suggest that every individual reader in a traditional news market constantly goes around checking all (or even some) news stories he reads. Rather, knowledge about the reputation of news outlets just is the kind of knowledge a citizen has. Michael Baurmann argues that we build trust-networks by observing the trust-relations of the people we trust already: if I trust my parents, and they trust my teacher, then I will trust my teacher, too.[83] This idea can be easily extended to include news outlets: a child might read the newspaper her parents read. Knowledge about the reputations of different news outlets is then embedded in one's general upbringing, and by serving as a proxy for truth this knowledge allows a market for news to function.

### 3. Truth heuristics on social media

But traditional markets for news are no more. News consumption increasingly takes place through social media.[84] And on social media, some problems arise. First, the number of news outlets active on social media is enormous. The reader is no longer constrained to three local papers and a dozen national ones: there are those, and then all other local papers, and then all national papers of all other countries, plus an almost endless collection of online news outlets. And this does not merely go for newspapers: the same holds true for tv and radio. Secondly, news on social media is not subscription-based: it typically spreads one article or video at a time. Thus a news consumer's social media feed is quickly filled with an array of articles from different sources from all over the planet. Heuristics like reputation simply do not work in such a sea of news, because the news consumer is extremely likely to come across a report from an unknown source, and due to the frequency with which this happens, the consumer cannot reasonably be expected to check the reputation of every source individually by tracking the truth of their exoteric statements.

This is a perfect example of the context-sensitivity of reliability which I discussed in chapter three. While the traditional news market was situated in an epistemic environment that allowed consumers to pick out reliable sources with relative accuracy – and thereby to navigate the division of epistemic labour – the new environment that

[83] Baurmann, "Fundamentalism and Epistemic Authority," 61.

[84] Ro'ee Levy, "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment," *American Economic Review* 111, no. 3 (March 1, 2021): 831–32, https://doi.org/10.1257/aer.20191777; Annika Bergström and Maria Jervelycke Belfrage, "News in Social Media: Incidental Consumption and the Role of Opinion Leaders," *Digital Journalism* 6, no. 5 (May 28, 2018): 583–84, https://doi.org/10.1080/21670811.2018.1423625.

social media offer renders the most important heuristics unreliable or insufficient. Thus the question emerges: which heuristic for truth-telling remains to the reader?

A likely candidate is simple pattern recognition. Dan Kahan argues that this is the sort of heuristic which laypeople use to recognise valid sources of science. Laypeople cannot recognise valid sources of science by attending to the content of what these sources say: the subject matter is simply too complex. Instead, they learn to recognise "the *signifiers* of validity implicit in *informal, everyday social processes*":[85] they learn to recognise *experts* rather than *truth*. A similar process can occur when news consumers attempt to recognise valid sources of news. Remember that news is the kind of thing which is produced by the journalistic process. If all news stems from the same kind of process, then it is likely to have some signifying elements. There is a certain *look* and *feel* to a news article – or video, or radio programme – that makes it distinct from other kinds of content.[86] In a pinch, and absent stronger heuristics, news consumers might turn to these signifying elements to establish whether a piece of content truly is news.

### 4. Deceptive news and fake news

I defined news as the product of journalistic practice, a defining characteristic of which is its dedication to truth. I now want to juxtapose this notion of news with something I will call "deceptive news". Deceptive news is content that is outwardly similar to news, but that is not the product of the journalistic process. Deceptive news is therefore likely to be false. The concept is relatively vaguely defined, and is therefore better conceived of as a spectrum than as a singular category with clearly defined boundaries: news satire is somewhere on the deceptive news spectrum, as well as some other types of fictional accounts. The notion of deceptive news is close to the notion of fake news, for which 'being similar to news' is a widely agreed upon requirement.[87] The two notions diverge, however, in the role intentions play. An oft-heard requirement for

[85] Dan M. Kahan, "On the Sources of Ordinary Science Knowledge and Extraordinary Science Ignorance," in *Oxford Handbook on the Science of Science Communication*, ed. Kathleen Hall Jamieson, Dan M. Kahan, and Dietram A. Scheufele, vol. 1 (Oxford University Press, 2017), https://doi.org/10.1093/oxfordhb/9780190497620.013.4.

[86] Edson C. Tandoc, Zheng Wei Lim, and Richard Ling, "Defining 'Fake News': A Typology of Scholarly Definitions," *Digital Journalism* 6, no. 2 (February 7, 2018): 11, https://doi.org/10.1080/21670811.2017.1360143.

[87] See, for example, Tandoc, Lim, and Ling, 147; Mathiesen, "Fake News and the Limits of Freedom of Speech," 167; Nikil Mukerji, "What Is Fake News?," *Ergo, an Open Access Journal of Philosophy* 5, no. 20201214 (December 11, 2018): 929, https://doi.org/10.3998/ergo.12405314.0005.035.

fake news is that it is *intentionally* deceptive. Mukerji Nikil, for example, argues that there must be an intention on the part of the producer of fake news to deceive the audience into believing that the report is an instance of real news.[88] There is no such requirement for deceptive news: it suffices for deceptive news to be similar to real news, regardless of whether those who produced it intended to deceive anyone.

Fake news is a subcategory of deceptive news: I define fake news as *intentionally deceptive news*. By this I mean the following:

*Fake news is content that:*

*1) is outwardly similar to news;*

*2) is not a product of a process whose first obligation is to the truth;*

*3) was created with the intention to deceive its audience into believing that it is news.*

This definition aligns closely with other philosophical definitions of fake news, such as Nikil Mukerji's and Regina Rini's.[89] The main difference with Mukerji's definition is his stipulation that claims must be asserted, and not merely implied. I do not disagree with this stipulation, but think that it is less obvious to stipulate in the present context due to the way I distribute my conditions: it is hard to see how intentionally deceptive news could fail to make any assertions at all. The main difference with Rini's definition is her stipulation that the story is known by its creators to be significantly false: on this point I align closer with Mukerji, who argues that it is enough that a creator of fake news simply not care about the truth of the story – which means that it might be accidentally true. Since I do stipulate that the story is not the product of a process whose first obligation is to the truth, however, the creator of the story does know that the story is likely to be false.

### 5. The rise of deception

Deceptive news is not a new phenomenon, and also not inherently problematic: it only becomes problematic when consumers mistake it for real news. This is problematic both economically and epistemically. It is problematic economically because in such a case deceptive news competes unfairly with real news, to the detriment of the latter. It is problematic epistemically because it negatively affects the reliability of the news market: if a consumer mistakenly takes a piece of content for news, then that consumer will expect

[88] Mukerji, "What Is Fake News?," 935.

[89] For further reference, see Mukerji, 929–30; Rini, "Fake News and Partisan Epistemology," E-45.

the creators of that content to have met the epistemic standards of journalism. But this expectation is mistaken, and therefore likely to lead to improperly held beliefs.

Whether consumers will often mistake deceptive news for real news depends on their ability to accurately recognise real news. The switch to simple pattern recognition as the main truth-heuristic, induced by the move of the news market to social media services, hinders this ability: since deceptive news looks much like real news, pattern recognition is an especially bad tool for those who wish to distinguish the two. To the extent that consumers have to rely on simple pattern recognition, they will be more likely to mistake deceptive news for real news – rendering the news market less adequate as a social-epistemic structure.

But this is just the beginning. Remember that social media are firmly situated in the digital attention economy. Not only social media corporations earn their money from human attention: the same goes for any other website which generates revenue from advertisements. The production of attention-generating content is a potentially lucrative business for any online actor. Jaster and Lanius argue that a great means to attract attention – in the terminology which I introduced in chapter two, a great hook – is simple, negative, and extraordinary news: [90] a news article claiming that the prime minister is involved in a corruption scandal is simply more interesting than an article about the qualities of subsections d through h of article 2 of a concept version of a new transatlantic trade deal. But traditional news outlets, with their reverence for truth, can only produce such attention-grabbing news every now and then, for they are dependent on what actually happens in the world.

This opens up a great business opportunity for the epistemically unscrupulous: the incentive structure of the attention economy combines with consumers' inability to distinguish real from deceptive news to open the door to actors that take advantage of the situation by producing content that imitates the look and feel of proper journalism, without letting themselves be constrained by requirements of truthfulness. Producing such fake news – for that is what it is – is much cheaper than producing real news: proper journalism is an arduous and costly process, but a producer of deceptive news merely needs a computer and an active imagination. Due to this strong incentive structure, we

[90] Romy Jaster and David Lanius, "Schlechte Nachrichten: >>Fake News<< in Politik Und Öffentlichkeit," in *Fake News Und Desinformation: Herausforderungen Für Die Vernetzte Gesellschaft Und Die Empirische Forschung*, ed. Ralf Hohlfeld et al. (Nomos Verlagsgesellschaft mbH & Co. KG, 2020), 248, https://doi.org/10.5771/9783748901334.

can expect the online news market to be flooded with fake news. This pushes the epistemic problems to ever greater heights by increasing the supply of deceptive news, which renders simple pattern recognition an even less reliable heuristic.

Thus we see that the ubiquity of fake news is a consequence of a problematic situation in which consumers have become unable to distinguish real from deceptive news. It is this problematic situation, which is essentially due to the deficient epistemic environment that social media services offer, which both spawns fake news and allows it to play its problematic role. But this role could be played by any kind of deceptive news, and in order to address the problem of fake news it is most important to address the root problem of a deficient epistemic environment – which itself is caused by the move from a traditional news market to a social media news market.

Because fake news will often offer more successful hooks than real news can, it will perform well in the eyes of attention-maximising recommender systems: news of the fake variety will do a good job at fulfilling the system's ODS. Noticing the success of these stories, the system will put the spotlight on them, using its curatorial power to cause them to reach even more users. These users, too, lack the tools to recognise these stories as fake, and thus the cycle continues. This could explain the incredible success of fake news in the run-up to the 2016 presidential election, while keeping at bay the conclusion that this success can only be explained by assuming that news consumers have lost their preference for truth: rather, it is the impoverished epistemic environment that social media services offer that renders its user incapable of correctly navigating the division of epistemic labour.

## 6. Conclusion: chasing the wrong preferences

In chapter three, I said that a good epistemic environment allows people to navigate the division of epistemic labour in order to acquire the information they need. In the present chapter, I have argued that the traditional news market, regulated by the reputation heuristic, is such an epistemically beneficial social structure. Journalism is a social practice that puts stringent epistemic demands on journalists, and partitioning this social practice into different news outlets allows the reputation heuristic to help non-journalists identify reports which meet these demands. Thus the traditional market for news increases general knowledge possession.

The move from the traditional market towards a market based on social media, however, throws several wrenches into this process. Most importantly, the great influx of news outlets renders the reputation heuristic less effective, causing people to rely on simple pattern recognition. But simple pattern recognition is easily fooled by deceptive news: content that looks like news, but that is not a product of the journalistic process. This already reduces the epistemic value of the news market, but the buck does not stop there: attention-economic incentives combine with consumers' inability to recognise real news in a toxic way to incentivise the production of fake news. This further changes the ratio of real to deceptive news in the latter's favour, rendering simple pattern recognition even less reliable.

In such an epistemically detrimental context, a recommender system that manages to track user preferences would be a blessing: such a system would prioritise real news over the fake variety, thereby providing counterweight to the epistemically vicious cycle. Recommender systems as they exist today, however, do no such thing. Indeed, an ODS that prioritises attention maximisation is likely to cause the system to propel fake news to even greater heights, because it performs so well in attention-economic terms. Rather than tracking and supporting user preferences, the system thus teams up with those producing fake news, rendering users even *less* capable of satisfying their preference for true news.

Recommender systems do not only fail to track user preferences: their failure to do so also renders the already epistemically hostile news-environment on social media even more vicious. This delivers two blows to the idea that social media are transparent. First of all, it undercuts the argument-from-preference-tracking by showing the premise that recommender systems track preferences to be mistaken. Secondly, the idea that social media services are transparent is challenged directly. Remember that customers in the traditional news market were relatively capable of finding trustworthy sources of accurate news. It is due to the move to social media that customers become less capable in this respect, which in turn causes the situation to deteriorate even further. All along, the social epistemic practice called "journalism" keeps operating the same way: it is only due to the intervention of social media services that citizens are not able to locate the products of this practice anymore. Thus the epistemic deterioration is entirely due to the choice on the side of the social media service to wield its curatorial power in a specific way. If the service chose to only propagate stories from trustworthy sources – as news outlets do in a traditional market – then the reliability of the news market would be salvaged. The choice not to do so is thus epistemically significant, giving us a first reason to believe that social media services are opaque.

## Chapter Five
## The Line: Rabbit Holes

When Caleb Cain dropped out of college, he had no conscious intention to start spending his days on YouTube. But stuck in his grandparents' house in rural Virginia, the videos on the platform offered solace and a sense of connection which he was lacking. At first he watched self-help videos, trying to lift himself out of his depression, to overcome his anxieties and give direction to his life. But after not too long he was watching various other kinds of content: videos about politics and philosophy led him to fringe scientific theories, and the more he watched the more he felt himself drifting away from everything he had thus far believed in. Within a short time, his liberal starting points had disappeared from his purview completely. In their place had come a bundle of racist, sexist and homophobic beliefs: an ideological package commonly associated with the term "alt right".

In his own words, Caleb "fell down the alt-right rabbit hole".[91] The phrase is well-known, and even the experience of going down a rabbit hole is familiar to most users of the internet: who has not navigated to Wikipedia to look up information about a specific topic, only to find themselves an hour later scrounging around in the most obscure corners of the world's largest online encyclopaedia? But rabbit holes lost their innocence in the second decade of the millennium, when stories like Caleb's started popping up. Somehow, it seemed, especially powerful rabbit holes were forming on YouTube that not only sucked people in with relentless force, but also ensued to radicalise them. Today, the term "rabbit hole" is often mentioned in connection with topics like the extreme right, radicalisation, and terrorism.[92]

If the YouTube is indeed to blame for the radicalising effects of its rabbit holes, then this might serve as another reason to doubt the transparency of social media services. In the present chapter, I mount an argument to that effect. Although rabbit holes play a prominent role in public debate about social media and specifically YouTube – the New

[91] The description of Caleb's story is based on his own, widely-cited YouTube video: Caleb Cain, *My Descent into the Alt-Right Pipeline*, 2019, https://www.youtube.com/watch?v=sfLa64_zLrU.

[92] Mark Alfano et al., "Technologically Scaffolded Atypical Cognition: The Case of YouTube's Recommender System," *Synthese*, June 9, 2020, https://doi.org/10.1007/s11229-020-02724-x; Derek O'Callaghan et al., "Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems," *Social Science Computer Review* 33, no. 4 (August 2015): 459–78, https://doi.org/10.1177/0894439314555329.

York Times even made an eight-part podcast series about them[93] – a good scientific or philosophical account does not yet exist. The aims of this chapter are therefore twofold. I start by constructing an account of what it means to go down a rabbit hole in general: what is this activity that sucks us in, and manages to direct our attention towards often quite unimportant facts? With this account in hand, I then consider what happens when a powerful attention-maximising recommender system is introduced into this process. It will become clear that, although going down a rabbit hole is in itself not an epistemically bad process, the epistemic carelessness that characterises social media recommender systems can turn this process sour, and indeed cause a gradual slope towards increasingly extreme content.

## 1. Down the rabbit hole

> *Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge. In another moment down went Alice after it, never once considering how in the world she was to get out again. The rabbit-hole went straight on like a tunnel for some way, and then dipped suddenly down, so suddenly that Alice had not a moment to think about stopping herself before she found herself falling down a very deep well.*[94]

Lacking a pre-existing account of rabbit holes on the internet, I will start at the very beginning: the term itself, and the way it is used in everyday discourse. The term "rabbit hole" is presumably a reference to this first chapter of Lewis Carroll's book *Alice's Adventures in Wonderland*. Alice's fall down the rabbit hole, which – at that point still unbeknownst to her – will transport her to Wonderland, is not an unpleasant experience. In fact, it is quite enjoyable: she falls so slowly that she has the time to look around, allowing her curiosity free roam as she inspects the shelves that line the burrow.

Compare Alice's experience to the following descriptions of rabbit holes on the internet, pulled from the crowd-sourced online *Urban Dictionary*:

[93] Kevin Roose, "Rabbit Hole," accessed June 1, 2021, https://www.nytimes.com/column/rabbit-hole.
[94] Lewis Carroll, *Alice's Adventures in Wonderland* (Boston: Lee and Shepard, 1869), 2–3.

> *To go down a never ending tunnel with many twists and turns on the internet, never truly arriving at a final destination, yet just finding more tunnels. Clicking one link, then finding another on that page, then clicking another link on that page, which gives you the idea to search for something, and the process repeats.*[95]

> *When you go online to search something and end up two hours later researching an obscure topic that has nothing to do with what you set out to look up.*[96]

There are, I think, three qualities which characterise the experiences of both Alice and internet users. First, there is a distinct sense of searching, investigating or researching something: a fall down a rabbit hole is driven by curiosity. Alice is drawn into the rabbit hole by her curiosity about a waistcoat-wearing watch-carrying white rabbit; the internet user goes online with the intention to "search something". Second, the object of investigation changes quickly and often: falling down the rabbit hole, Alice gets distracted inspecting the shelves, picking up jars and carefully putting them back, forgetting entirely the rabbit she is chasing; the internet user clicks from page to page, and each new page inspires him to research a brand new topic. Third, the way down the rabbit hole is a fall, not a climb: Alice must accept that the fall will only end once she hits the ground; the internet user "ends up" hours after starting his search researching an obscure topic – as if he lost track of time, only zoning back in at a much later point.

If we take these three qualities – an investigative attitude driven by curiosity, oft-changing objects of investigation, and a sense of ease accompanied by a loss of sense of time – as characterising the fall down an online rabbit hole, then it becomes clear that we should conceive of "falling down a rabbit hole" as an *activity*, rather than conceiving of rabbit holes as objects or structures that one happens to fall down. The online rabbit hole does not exist independently of the user falling down it: what makes the links which the user clicks parts of a rabbit hole is the fact that the user clicks them whilst falling. The rabbit hole is the investigative path which the internet user happens to take, and it is a rabbit hole because the user takes that path.

[95] "Urban Dictionary: Rabbit Hole," Urban Dictionary, August 19, 2013, https://www.urbandictionary.com/define.php?term=Rabbit%20Hole.

[96] "Urban Dictionary: Rabbit Hole," Urban Dictionary, June 17, 2015, https://www.urbandictionary.com/define.php?term=Rabbit%20Hole.

This definition allows that one could fall down a rabbit hole in *any* kind of environment that facilitates research: the definition is not restricted to the internet. One might fall down a rabbit hole in a library, allowing references in one book to guide one to the next book, on and on across the aisles. Yet not *all* investigative activities in a library count as a fall down a rabbit hole. Consider a person who spends an afternoon browsing the shelves, picking up books, reading a page or two and putting them back again. If there is no reason why this person picks up *these* books in particular, if there is nothing that *connects* the different books this person peruses, then intuitively we should not describe that person as falling down a rabbit hole. A fall down a rabbit hole is characterised by oft-changing objects of investigation, but these changes may not be entirely random: in order to qualify as falling down a rabbit hole, a person's interest in topic B must be sparked in the process of researching topic A; interest in C must be sparked while researching B; et cetera. Although the rabbit hole may have many unexpected twists and turns, making it seem very random indeed, it must in essence remain a path.

This is presumably why a fall down a rabbit hole is so much more likely to occur online than in the analogue world, for it is *so easy* to follow a path online. Allowing oneself to be sent down a rabbit hole in a library seems a rather cumbersome process, for one is continuously distracted by all kinds of activities that have nothing to do with the topics one is actually curious about: one must find the right book, then find the right page, set the book back again, locate the next book, find it, find the right page, et cetera. But when falling down a rabbit hole on Wikipedia – sometimes aptly called a "wikihole"[97] – none of these distractions are there to pull one out of the flow: the link to topic B is embedded in the text about topic A, and all one need to do is click on it to be transported to one's next pursuit.

## 2. The role of curiosity

We gain a deeper understanding of rabbit holes by taking a closer look at the mental state that drives a fall down a rabbit hole: curiosity. In this context, Frederick Schmitt and Reza Lahroodi's account of curiosity as an epistemic virtue is extremely valuable. Schmitt and Lahroodi (S&L) argue that an occurrent state of curiosity (as opposed to a disposition to be curious, or curiosity as a character trait) can be best understood as a motivationally

[97] "Wiki Rabbit Hole," in *Wikipedia*, April 2, 2021, https://en.wikipedia.org/w/index.php?title=Wiki_rabbit_hole&oldid=1015623747.

original desire to know a topic combined with one's attention being drawn to the topic.[98] That the desire is motivationally original means that there is no extraneous reason to know the topic: one does not wish to know in order to do something with that knowledge, or even because one wishes to have a great amount of knowledge in general; one simply wishes to know for the sake of knowing. The attention which is directed at the topic is causally interconnected with the motivationally original desire to know: the desire arises because one's attention is first drawn to the topic, and the attention for the topic is then sustained by the desire to know.[99]

S&L situate their account in the field of virtue epistemology.[100] Although this field is somewhat removed from the reliabilist and epistemic consequentialist framework which I have been working with, the value of curiosity is easily identified within my framework as well. S&L write: "Curiosity has the value of fixing our desire to know topics into which we would not otherwise be motivated to inquire, thereby making us attend to them more closely than we would otherwise."[101] This renders curiosity clearly valuable in an epistemic consequentialist framework as well: to the extent that the desire to know translates into actual knowledge, curiosity increases knowledge-possession. And since curiosity is a desire to *know*, we can expect it to lead to knowledge just in those cases where the subject understands the demands of epistemic propriety: if the subject knows what "knowing" means, then she will aim to fulfil her epistemic responsibilities diligently in her quest for knowledge.

S&L identify two additional features of curiosity that both contribute to its epistemic value. Firstly, curiosity "does not generally depend for its choice of topics on our having prior practical or epistemic interests in that topic."[102] Due to this, curiosity causes us to acquire a broad knowledge base. Secondly, curiosity is *tenacious*: "for a typical state of curiosity whether *p*, one has more than a desire to know whether *p*; one is also disposed to be curious about issues related to *p*."[103] Tenacious curiosity is valuable because "a tenacious state of curiosity will eventuate in a larger body of knowledge related to the

[98] Frederick F. Schmitt and Reza Lahroodi, "The Epistemic Value of Curiosity," *Educational Theory* 58, no. 2 (May 2008): 128, https://doi.org/10.1111/j.1741-5446.2008.00281.x.

[99] Schmitt and Lahroodi, 129.

[100] Schmitt and Lahroodi, 126. Somewhat frustratingly, S&L fail to specify in *which* virtue epistemology their account is supposed to fit.

[101] Schmitt and Lahroodi, 133.

[102] Schmitt and Lahroodi, 140.

[103] Schmitt and Lahroodi, 137.

topic of curiosity than a nontenacious state of curiosity will (…) In short, it will lead to *deeper* knowledge of the topics of curiosity than nontenacious states do."[104]

Being a process driven by curiosity, we can expect a fall down a rabbit hole to generally be an epistemically valuable activity as long as it occurs in an environment where one's cognitive processes remain reliable. Indeed, in such an environment a fall down a rabbit hole will not only lead to more knowledge: it will both broaden and deepen one's knowledge base.

The tenacity of curiosity provides a satisfying explanation of the path-like quality of rabbit holes. Assuming that one's initial quest is induced by curiosity about a certain topic, the tenacity of curiosity leads us to expect the object of curiosity to switch to a related topic the moment one's initial curiosity is satisfied. Thus it is because curiosity is regenerative in this way that the investigative activity of looking up a single topic on Wikipedia becomes iterative, changing from an investigation of one topic into an investigation of another, over and over.

Just one feature of the activity of falling down a rabbit hole remains unexplained: what makes this process so enjoyable or easy – and what does the loss of sense of time have to do with it? It seems completely unrelated to the other two criteria, and yet it is an essential part of the experience of falling down a rabbit hole. To answer this question, we will have to go down a short a rabbit hole of our own: a short foray into the world of Las Vegas' slot machines will not only teach us about rabbit holes, but also bring us one step closer to understanding the peculiar power of YouTube's recommender system.

### 3. What happened in Vegas?

Over the course of almost two decades, Natasha Dow Schüll conducted anthropological research into the gambling culture in Las Vegas. Her studies focused on gambling addiction, and specifically addiction to slot machines. In her book *Addiction by Design,* Schüll first describes how slot machines are designed to induce so-called *continuous gaming productivity*, which "involves three interlinked operations (…): *accelerating* play, *extending* its duration, and *increasing* the total amount spent."[105] Schüll then ensues to speak to those whose lives are most affected by these technologies:

[104] Schmitt and Lahroodi, 139.
[105] Natasha Dow Schüll, *Addiction by Design: Machine Gambling in Las Vegas* (Princeton, NJ: Princeton University Press, 2012), 52.

gamblers themselves. Schüll’s findings are of particular interest because in the gambling industry, we find a rare example of a mature industry that employs advanced interactive technologies in order to capture and extend human attention: in some ways, the business model of the gambling industry is very close to that of social media services. In this section, I trace Schüll’s account of gambling addiction, and show that the theoretical apparatus which Schüll employs can be used to gain a deeper understanding of rabbit holes.

One of Schüll’s most remarkable findings is that those who deal with gambling addiction are usually not “in it to win it”. Indeed, the animations and noises that come with winning a jackpot are often experienced as annoying interruptions of play, and players do everything in their power (such as smashing the play-button, or inserting more coins) to return to the game itself when these interruptions occur.[106] In the following conversation, Mollie, one of Schüll’s interviewees, describes what she gets from gambling – an experience that is echoed by other gamblers throughout the book:

> *“The thing people never understand is that* I’m not playing to win.”
>
> *Why, then, does she play? “To keep playing – to stay in that machine zone where nothing else matters.”*
>
> *I ask Mollie to describe the machine zone. (…) “It’s like being in the eye of a storm, is how I’d describe it. Your vision is clear on the machine in front of you but the whole world is spinning around you, and you can’t really hear anything. You aren’t really there – you’re with the machine and that’s all you’re with.”*[107]

The zone – the term is used throughout the book by gamblers and game designers alike – is an almost dissociative state: it is “like a magnet, it just pulls you in and holds you there.”[108] The state is rewarding not only to gamblers: a machine that induces the zone in gamblers is also the holy grail of any gambling machine designer pursuing continuous gaming productivity, for an important aspect of the zone is that the gambler loses track of time, trapping her behind the machine for as long as multiple days in a row.[109]

[106] Schüll, 224.
[107] Schüll, 2.
[108] Schüll, 19.
[109] Schüll, 179.

Schüll explains the machine zone by means of the psychological concept “flow”. This concept was introduced by Mihaly Csikszentmihalyi as a result of research into the subjective phenomenology of autotelic activities: activities that are rewarding in and of themselves (also known as intrinsically motivated activities).[110] Interviewing people who participated in activities which they experienced as autotelic, ranging from chess players to rock climbers, Csikszentmihalyi found that the subjective state which these people reported achieving whilst participating in the activities were “remarkably similar across play and work settings”.[111] It is this state which Csikszentmihalyi termed “flow”. The state has an array of characteristics that together have a desubjectifying effect: Schüll describes flow-states as “states of absorption in which attention is so narrowly focused on an activity that a sense of time fades, along with the troubles and concerns of day-to-day life”.[112]

Csikszentmihalyi identifies four conditions for flow:

1. There must be “clear proximal goals”,[113] or in Schüll’s words, “each moment of the activity must have a little goal”;[114]
2. It must be clear how these goals are to be achieved;[115]
3. There must be “immediate feedback about the progress that is being made”;[116]
4. Perhaps most importantly, the “perceived challenges, or opportunities for action”[117] must “stretch (neither overmatching nor underutilizing) existing skills”:[118] the activity calls upon our skills in such a way that it hits the sweet spot between strenuousness, which would cause anxiety, and ease, which would cause boredom.

Schüll argues that machine gambling meets each of these conditions – or at least, the machines emulate the experience of meeting these conditions. And indeed, machine gamblers’ descriptions of the “zone” line up almost perfectly with the phenomenology of

[110] Jeanne Nakamura and Mihaly Csikszentmihalyi, “The Concept of Flow,” in *Handbook of Positive Psychology*, ed. C.R. Snyder and S.J. Lopez (Oxford: Oxford University Press, 2005), 89, https://doi.org/10.1007/978-94-017-9088-8_16.
[111] Nakamura and Csikszentmihalyi, 89.
[112] Schüll, *Addiction by Design*, 166.
[113] Nakamura and Csikszentmihalyi, “The Concept of Flow,” 90.
[114] Schüll, *Addiction by Design*, 166.
[115] Schüll, 166.
[116] Nakamura and Csikszentmihalyi, “The Concept of Flow,” 90.
[117] Nakamura and Csikszentmihalyi, 90.
[118] Nakamura and Csikszentmihalyi, 90.

flow: "Gamblers "forget themselves" and feel carried forward by a choreography not of their own making; much like mountain climbers who describe merging with the rocks they climb, or dancers who report feeling "danced" by music, they feel "played by the machine."[119] The only difference that Schüll notes is that while Csikszentmihalyi's interviewees typically describe flow as a positive state, life-affirming, restoring and enriching, machine gamblers feel trapped by their zone, depleted, and deprived of their autonomy.[120]

Schüll's description of technology that generates this 'dark side of flow' has drawn attention, especially among those studying games and game-like technologies like social media. The gambling industry is not the only one to have hit upon a formula that induces a zone-like (or flow-like) state: similar formulae have been observed in simple mobile games like Tetris[121] and Flappy Bird,[122] and – notably for us – in the algorithms of social media.[123] Outside of academic print, Schüll has introduced the term "ludic loop" for the characteristics that these technologies seem to share with advanced gambling machines.[124] Ludic loops are structures which interactive technologies can possess that, when successful, serve to keep users in the machine zone. Schüll identifies four key components of ludic loops:[125]

1. *Solitude*: the interaction with the technology excludes other people, which facilitates continued interaction because there are no stopping cues originating from outside of the interaction;
2. *Immediate feedback*: because there is no delay between user action and machine reaction, there are no natural points at which to stop or pause the activity;

[119] Schüll, *Addiction by Design*, 167.
[120] Schüll, 167.
[121] Paul Gapper, "A Satisfying Plot 5: Questions and Puzzles," *Paulgapper* (blog), June 1, 2014, https://paulgapper.wordpress.com/tag/ludic-loop/.
[122] Sam Akester, "Ludic Loop," *Sam Akester* (blog), November 11, 2015, https://sammehreviews.wordpress.com/2015/11/11/ludic-loop/.
[123] Xing Lu, Zhicong Lu, and Changqing Liu, "Exploring TikTok Use and Non-Use Practices and Experiences in China," in *Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing*, ed. Gabriele Meiselwitz, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2020), 57, https://doi.org/10.1007/978-3-030-49576-3_5.
[124] Natasha Dow Schüll, NPR All Things Considered, interview by Arun Rath, Radio, June 7, 2014.
[125] Regrettably, the record of the lecture in which Schüll introduced these characteristics is no longer accessible. She has privately confirmed their importance, however. For a more elaborate account, see B David Zarley, "Inside the Fight to Make Tech More Humane," A Beautiful Perspective, March 2, 2018, https://abeautifulperspective.com/2018/03/inside-the-fight-to-make-tech-more-humane/.

3. *Random rewards*: rewards that happen occasionally in an unpredictable pattern are known to behavioural psychology to cause people to compulsively continue (gambling) behaviour;[126]
4. *Non-resolution*: just like the solitude requirement ensures that there are no stopping cues originating from outside of the interaction, there are also no stopping cues originating from inside the interaction: the process could go on and on.

Note that each of these components serves to keep the user engaged: solitude, immediate feedback, random rewards, and non-resolution are strategies to trap the user in an activity. Very little is said, however, about the nature of the activity. This is problematic because it is easy to think of counterexamples: many activities seem to meet the criteria for a ludic loop, yet it is unlikely that they will all successfully trap the user in a zone-like state. Think, for example, of a simple machine consisting of a single button hooked to a screen. The player, alone in a room, is told to press the button in any rhythm and at any speed he desires. Every so often – and at random intervals – a press of the button makes the screen light up, broadcasting the message "You've won!" Now, empirical research might be needed to say this with certainty, but it seems a safe guess to say that this 'game' would fail to hold anybody's attention for long – even though it meets all the criteria for a ludic loop.

What distinguishes my counterexample from games like Flappy Bird or Tetris? The difference is that successful games do not only incorporate the components of a ludic loop, but also meet Csikszentmihalyi's criteria for flow-inducing activities. The addictive quality of these games can thus be understood as a toxic interplay between flow-inducing characteristics on the one hand, and ludic loop characteristics on the other: while the flow-inducing characteristics induce a flow-like state of extreme attentiveness, the ludic loop characteristics trap the user in that state by eliminating stopping cues as well as natural moments to decide to do something else. Together, these two elements cause the machine zone: an endless state of suspension in a world of play.

[126] Richard J. E. James, Claire O'Malley, and Richard J. Tunney, "Understanding the Psychology of Mobile Gambling: A Behavioural Synthesis," *British Journal of Psychology* 108, no. 3 (2017): 613, https://doi.org/10.1111/bjop.12226.

### 4. Alice in Vegas

As we emerge again from Las Vegas, the central thesis of this chapter now stands clearly before us. I submit that to fall down a rabbit hole is to engage in a flow-inducing ludic loop: the rabbit hole *is* the machine zone. In this section I argue for this thesis by evaluating the activity of falling down a rabbit hole on Wikipedia: I consider the phenomenology of the activity and tick the boxes for ludic loops and flow activities. Having thus firmly established that falling down a rabbit hole can be seen as a machine zone inducing activity, I consider in section 4 what happens when a powerful attention-maximising recommender system is introduced into the process, arguing that YouTube's choice of ODS can have distinct and detrimental epistemic effects.

Let's start with the phenomenology of Wikipedia rabbit holes. For ease of reference, I will call our internet user who falls down a rabbit hole "Alice". Note how closely our descriptions of rabbit holes resemble the phenomenology of flow states. In both, the subject performs an activity, but there is a distinct sense in which the activity simply 'happens': the dancer is danced by the music, the climber merges with the rock, the machine plays the player – and Alice falls down the rabbit hole. Connecting these descriptions is a loss of self-awareness on the side of the subject: although seen from the outside as a matter of a subject participating in an activity, the subjective experience is one of merging with the activity, so that it does not feel to the subject as if she is actively deciding to do anything. This causes a loss of sense of time: the dancer and the climber simply keep going until the end of the song or the top of the mountain; the player sits at her machine for entire days; and Alice 'zones back in' hours later, having drifted off to a deep corner of Wikipedia.

On top of being phenomenologically similar to flow experiences, falling down a rabbit hole meets all the criteria for being a ludic loop. To start, it is an activity best performed alone: browsing Wikipedia in tandem is difficult. Secondly, there is immediate feedback: a click on a link immediately transports Alice to a next page – provided that her internet connection is good enough. Indeed, the prospect of falling down a rabbit hole on a bad internet connection sounds, if anything, frustrating: having to wait long periods for the next page to load seems likely to pull Alice out of the rabbit hole after not too long. It was for similar reasons that it seemed unlikely that a true rabbit hole can be achieved by scouring through a library. Third, Alice is likely to encounter random rewards as she navigates Wikipedia: not all pages she visits will be equally interesting, but every so often she bumps into a little goldmine. And fourth, the rabbit hole is a non-resolving activity: since the premise of the rabbit hole is that Alice does not look for anything in particular, but rather lets herself be guided by the information she encounters, the activity can in principle go on and on.

To see how falling down a rabbit hole can also tick the boxes of flow-inducing activities, we need to return to the mental state behind a fall down a rabbit hole: curiosity. If we follow S&L in conceptualising curiosity as a motivationally original desire to know a topic, then the iterative investigative activities driven by curiosity that characterise rabbit holes emerge as autotelically structured: the process of allowing one's curiosity to be piqued and satisfying it, again and again, is entirely motivated by the activity itself. It is an interplay between desiring to know a topic for the sake of knowing, and getting to know that topic. This is an important point, since – as I noted in the previous section – flow is typically connected with autotelic activities.

Schmitt and Lahroodi's account helps us recognise two more characteristics of flow-inducing activities. First, the tenacity of curiosity ensures that there are clear proximal goals: Alice now has a desire to know *p*, now to know *q*. Secondly, it is clear how these goals are to be achieved: Alice can come to know *p* by reading and understanding the Wikipedia page about *p*, and similarly for *q*. But can we also recognise the third (clear immediate feedback about progress) and fourth (challenges that stretch skills) in the activity of falling down a rabbit hole?

I think we can, but in order to do so we must first see how the two requirements are connected. Note that it is difficult to tell whether our skills properly match a certain challenge without attempting to meet that challenge: a mathematical problem can seem very simple to solve, but turn out to be extremely difficult; an opposing sports team can seem a tough opponent, but prove easy to beat. Only when the mathematician engages the problem, things might start falling into place, and when our own team starts racking up points, it becomes clear that we are up to the challenge. Thus, clear and immediate feedback is necessary for flow because it serves as an indication that one's skills are a match for this particular challenge. It is for this reason that Nakamura and Csikszentmihalyi emphasise that "It is the subjective challenges and subjective skills, not objective ones, that influence the quality of a person's experience".[127] This is not to say

[127] Nakamura and Csikszentmihalyi, "The Concept of Flow," 91.

that it is impossible to tell objectively whether the skills are met by the challenge, but rather that the subjective experience of skills meeting challenges has priority over this objective measure. Given reliable feedback, objective and subjective will often map on to each other: an objectively bad rock climber will fall, which will give her the subjective experience that she is not up to the challenge. But at times the two might diverge because the available feedback is misleading – and in these cases the subjective experience will determine whether flow is sustained or not.

Falling down a rabbit hole is a distinctively epistemic activity. This means that the relevant capacities must be epistemic in nature, too: they are those capacities which are involved in the process of coming to know or understand[128] something by means of investigative activities. The question which now arises, is: what can we take as relevant feedback with regards to the question whether we are utilising our epistemic capacities correctly and effectively? The piano player hears beautiful music when she performs well, the rock climber glides smoothly up the rock. What kind of information is available to Alice?

C. Thi Nguyen argues that there is in fact very clear feedback available, for "the moment when we come to understand often has a particular feel to it – what some philosophers have called the "a-ha!" moment."[129] He calls these phenomenal states which are associated with understanding *(the sense of) clarity*. The sense of clarity subdivides into two parts: there are phenomenal states associated with coming to understand, and phenomenal states associated with having an understanding. When we come to understand, we have a feeling of 'things falling into place': "our system of thought changes and pieces of information that we could not accommodate before suddenly find a place".[130] Nguyen calls this phenomenon *cognitive epiphany*.[131] When we have an understanding, on the other hand, we experience *cognitive facility*: we can use our

[128] Schmitt and Lahroodi exclusively speak of curiosity as a desire to know, but a desire to know must sometimes include a desire to understand. If I am curious about the presents I will get for my birthday, this curiosity is most aptly described as a desire to know: I wish to know propositions of the form "I will get x for my birthday". But if I am curious why airplanes stay up, then simple propositional knowledge will not do: I will require an understanding of airplane mechanics before I can have propositional knowledge about the behaviour of airplanes. In these cases, curiosity can be better described as a desire to understand. I will follow Nguyen, "The Seductions of Clarity." In this regard, conceptualising "understanding" as a grasp of how facts connect: "Understanding is of a system; it involves grasping a structure and not just independent nodes." (p. 239)

[129] Nguyen, 228.

[130] Nguyen, 239–40.

[131] Nguyen, 240.

understanding to quickly generate predictions and further explanations.[132] Being clear phenomenal states which go hand in hand with understanding, both cognitive epiphany and cognitive facility can function as feedback for Alice. The sense of clarity can serve as an indication that one's curiosity for a specific topic has been satisfied – which opens the floor for the emergence of curiosity about a new topic.

Thus it is her own sense of clarity which tells Alice that she is using her epistemic capacities correctly and efficiently. Whether she has the subjective experience of skills stretching challenges depends both on Alice's epistemic capacities as well as on the topics she is researching: no uniform answer can be given to this question. Do note, however, that falling down a rabbit hole is a self-directed activity, and it is therefore to a certain extent up to Alice whether the challenges are a good match for her skills: if the topics she researches are so simple they bore her, she would be wise to click on more complicated-sounding links. We might even say that whether Alice goes down a rabbit hole is dependent exactly on whether she performs this task correctly: the rabbit hole emerges only when the internet user allows her curiosity to be piqued by those topics which require the right amount of skill to keep her engaged.

### 5. Alice Seduced

My focus thus far has been on rabbit holes on Wikipedia. Wikipedia was a good place to start because it is a rather passive service: although it offers the infrastructure that makes it possible to follow an investigative path, it does not amend this infrastructure very often. It does not matter who is exploring Wikipedia, or when, or where they are: entries of the encyclopaedia are interlinked in the same way for everyone. On top of this, Wikipedia's curatorial process has a distinctly epistemic character aimed at producing reliable content. Therefore, to the extent that this process is successful – and indications are that the process is fairly effective[133] – Wikipedia is a relatively safe environment, epistemically, for falling down a rabbit hole.

How different this is for social media services. Remember that most social media services use recommender systems to decide which content it shows the user next. In the case of rabbit holes, the best example is the video service YouTube, because the way it is

[132] Nguyen, 240.

[133] Jona Kräenbring et al., "Accuracy and Completeness of Drug Information in Wikipedia: A Comparison with Standard Textbooks of Pharmacology," *PloS One* 9, no. 9 (2014): e106930.

designed still affords the user a sense of agency. Next to the video which the user is currently watching, YouTube shows its – algorithmically generated – recommendations for the next video to watch in the so-called sidebar. This sidebar is infinite in principle, but without scrolling down the user can see approximately eight recommended videos. This setup is perfect for inducing rabbit holes, provided that the recommender system manages to recommend videos that would satisfy the curiosity which was piqued by the videos the user just watched. The element of choice between multiple videos makes this more likely.

We know from chapter four that YouTube has, at least in the past, used an ODS which prioritised attention maximisation. Just like the machine zone is the holy grail for gambling machine designers – as I argued in section three of the present chapter – so the rabbit hole is a holy grail for an attention-maximising recommender system, for the main defining feature of a fall down a rabbit hole – extended and extremely focused attention – is entirely in line with the ODS of such a system. It is therefore reasonable to expect that YouTube's recommender system will do anything in its power to facilitate falls down rabbit holes. But how can a recommender system achieve this?

Let me first state what might be obvious by now: a recommender system is not capable of "making" rabbit holes, any more than a city planner can decide which path a particular citizen will take to work. It is at the very core of the concept of falling down a rabbit hole that the user – I will refer to her again as "Alice" – has a degree of choice in the path she takes. It is Alice's curiosity that leads her this way rather than that.

What the recommender system *can* do, is change the lay-out of the city: it can change which paths Alice can choose from. If the assumptions on which a recommender system operates – most importantly, the assumption that Alice will react to content similarly to how people who showed sufficiently similar behaviour to Alice's did – are correct, then over time it will learn how a given video will pique Alice's curiosity: it will learn that when Alice watches video X, she will afterwards be curious about topics W, Y or Z. Moreover, given an array of videos about topics W, Y and Z, the recommender system will learn which of these videos is most likely to give Alice a sensation that it calls upon her epistemic capacities at exactly the right level to keep her in the flow, neither overstretching her skills nor underutilising them. In other words, the recommender system would learn which videos call forth the sense of clarity at *just* the right time to keep Alice engaged.

It is here that a severe epistemic danger arises, as a result of an interplay between what Nguyen has called the seductions of clarity, and Schmitt and Lahroodi's tenacity of curiosity. Remember that I defined clarity – following Nguyen – as the phenomenal state which is associated with understanding. But, as Nguyen argues, "(c)larity may often accompany genuine understanding, but it is by no means a perfect indicator that we do, in fact, genuinely understand."[134] If Nguyen is right, then clarity can be emulated: certain accounts or systems of thought can induce the sense of clarity without offering actual understanding. The emulation of clarity is a trick which can be employed by those who wish to epistemically manipulate their victims, because the sense of clarity serves as a thought-terminating heuristic: when we achieve clarity, we take it as a sign that we need not consider the matter further. Thus if an epistemic malefactor wishes to convince someone of the truth of a system of thought that would, on closer inspection, turn out to be deficient, then it is helpful if that system of thought triggers the sense of clarity very quickly: when it does, the victim will stop inquiring, and simply accept.[135] Nguyen argues that epistemically questionable structures like conspiracy theories and echo chambers often use this exact trick.[136]

This is a problematic state of affairs, because the YouTube recommender system cannot distinguish between videos that result in genuine understanding, and videos that merely induce a sense of clarity. It is, however, a fact that genuine understanding is extremely hard to achieve in many areas. Many topics studied by the sciences, for example, require year-long training in order to even approach understanding. If Alice is curious about such topics, then she does not have a chance of satisfying her curiosity by means of genuine understanding, merely by watching a YouTube video or two. Any attempt in that direction would likely severely overstretch her epistemic skills, which would cause flow-terminating anxiety, or simply result in failure, which would end the ludic loop. Thus, given the choice between a flow-terminating video aimed at genuine understanding, and a flow-sustaining video aimed at merely inducing the sense of clarity, we can expect the recommender system to recommend the latter, because it does not see the difference between the two except for the fact that one of them helps it achieve its goal: sustaining the fall down the rabbit hole.

[134] Nguyen, "The Seductions of Clarity," 232.
[135] Nguyen, 235–36.
[136] Nguyen, 243–44.

So, after watching a news report on YouTube about the storming of the United States capitol, Alice is not presented with a suggestion to watch a video about the sociology of polarisation, for that would terminate her fall down the rabbit hole; instead, she clicks on the recommended video about the deep state, an extreme right conspiracy theory which purports to explain why the United States government is (supposedly) dysfunctional. After watching this video, what will Alice do? The tenacity of curiosity ensures that curiosity about a given topic induces curiosity about topics related to that topic. The tenacity of Alice's curiosity about the deep state causes Alice to now be curious about topics related to the deep state. Perhaps the video about the deep state mentioned the supposed revelations of Q, protagonist of the conspiracy called QAnon. In that case, Alice might now be curious about Q, and will therefore likely be sustained in her fall down the rabbit hole if she watches a video about Q.

We see that the interplay between the seductions of clarity on the one hand and the tenacity of curiosity on the other can explain why rabbit holes on YouTube are associated with phenomena like extremism and conspiracy theories. The seductions of clarity first pull the internet user in the direction of extreme (simplistic) content, while the tenacity of curiosity serves to keep the user in that corner – or even draw her deeper into it. This is epistemically problematic because beliefs held due to the seductions of clarity are necessarily improperly held: the seductive clarity caused us to terminate our cognitive processes prematurely, thereby rendering them unreliable.

If this argumentation is correct, we should expect that YouTube's recommender system will show two distinct behaviours. First, it should at times have a tendency to recommend videos that induce a sense of clarity over genuine understanding – the most important example of such videos are conspiracy videos. Second, once it has shown a mere clarity-inducing video, it should keep recommending such videos. Empirical studies of the behaviour of YouTube's recommender system have shown these exact two behaviours.[137]

## 6. Distinctly epistemic curation

In rabbit holes we find another phenomenon that illustrates the opacity of social media services. Note that once again the recommender system does not track

[137] O'Callaghan et al., "Down the (White) Rabbit Hole"; Alfano et al., "Technologically Scaffolded Atypical Cognition."

preferences: if anything, the system aims to predict the user's occurrent states of curiosity. Thus we have uncovered a second reason to doubt the argument-from-preference-tracking which was introduced in chapter three. Moreover, the interventions of the recommender system once again reduce the reliability of our cognitive processes: while a fall down a rabbit hole is epistemically beneficial in a good environment, it can have detrimental consequences on a service like YouTube.

In both the deceptive news and the rabbit hole cases, the recommender system merely aims to maximise attention. In the first case, epistemic problems arise because some of the available content is deceptive (or even fake) news. Since such deceptive content is available, the recommender system sometimes recommends it, with epistemically undesirable consequences. This problem can thus be largely solved by only allowing real news outlets to share news stories on social media: no change to the recommender system is necessary.

The present case is different, however. The recommender system does not merely show a disregard for truth: it explicitly employs distinctly epistemic characteristics of certain content. The system's goal to maximise attention leads it to sometimes recommend certain content *exactly because* it is misleading. Alice's pursuit of truth about the storming of the capitol *should* have stranded in confusion about the sociology of polarisation, but by recommending seducingly clear accounts, the system managed to hold Alice's attention for a little while longer. Thus the negative effect that the YouTube recommender system has on the knowledge-conduciveness of the activity is not a mere unfortunate consequence of the fact that the recommender system operates in an untrustworthy environment: the epistemic goal of increasing knowledge possession and the attention-economic goal of maximising attention are at odds, and the seductions of clarity are explicitly enlisted in the recommender system's pursuit of attention.

In rabbit holes we have therefore found a first example of an explicitly *epistemic strategy* which a recommender system can use in its pursuit of attention, with epistemically detrimental consequences. The availability of such strategies to the recommender system shows that the curatorial power of social media services can indeed be characterised as epistemic: the recommender system can achieve its ends by manipulating the reliability of users' belief-forming processes. In the next chapter I consider a number of other epistemic strategies that are available to the recommender system: strategies that explicitly employ belief manipulation as a means to maximise attention.

### 7. Conclusion: is this Wonderland?

I have argued for two distinct but interdependent theses. The first is that a fall down a rabbit hole is a flow activity shielded by a ludic loop, inducing a state akin to what Natasha Dow Schüll termed the "machine zone": a state of utmost concentration without natural stopping points. But unlike the gambling activities for which Schüll uses the term, a fall down a rabbit hole is a distinctly epistemic activity, driven by iterative occurrent states of curiosity. Schüll describes the lure of the gambling ludic loop as a constant switching between certainty and uncertainty: "It's open, close, open, close";[138] a fall down a rabbit hole offers a similar dynamic, in the form of a constant switching between the uncertainty of curiosity and the resolution afforded by the sense of clarity.

Secondly, the importance of clarity in this process makes the fall down the rabbit hole vulnerable to epistemic manipulation. Often, it is easier to induce a mere *sense* of clarity than actual understanding. Nguyen compares clarity to a certain kind of culinary yumminess. Once upon a time, he says, we could trust our inclination towards this kind of yumminess to lead us toward nutritious food, for the one was a good heuristic for the other. "But our nutritive environment changed, especially when various corporate forces figured out our heuristics and tendencies and started to aggressively game them."[139] A very similar thing happens on platforms driven by algorithmic recommendations, most notably YouTube. Just like foods that are too yummy can send us into an eating frenzy, the recommender systems of these platforms aim to send us into an epistemic frenzy. In trying to induce this frenzy, the recommender system does not care whether it serves its users truths, half-truths or falsehoods: what matters is that the user experiences a sense of clarity.

In an epistemically benign environment, a fall down a rabbit hole is epistemically beneficial: our curiosity leads us into new and hitherto unexplored terrain, all the while enlarging our knowledge base. But this epistemically valuable behaviour starts producing vicious outcomes when curation of the epistemic environment is handed over from

[138] Douglas Heaven, "Engineered Compulsion: Why Candy Crush Is the Future of More than Games," *New Scientist* 222, no. 2971 (May 2014): 40, https://doi.org/10.1016/S0262-4079(14)61069-1.
[139] Nguyen, "The Seductions of Clarity," 250.

human, epistemically motivated actors (like Wikipedia's volunteers) to an attention-maximising, epistemically indifferent recommender system of the kind that YouTube employs.

The size of the problem is even more impressive when we consider that other feature of curiosity that Schmitt and Lahroodi identify: independence from our interests. For Schmitt and Lahroodi, this independence from our interests is another reason why curiosity is valuable: just like the tenacity of curiosity ensures an increased depth of our knowledge by inquiring more into topics related to the one we are now researching, so the independence from our interests ensures an increase in breadth of knowledge. But if my argumentation so far has been correct, then this feature will only increase the vicious effects of YouTube's recommender system, for it can cause people to be drawn into epistemically vicious rabbit holes who had no intention of going down them: an internet user need not even have a prior interest in politics in order to be seduced by the clarity of political conspiracy theories. And once seduced, who knows where Alice will end up?

*Down, down, down. Would the fall* never *come to an end?*[140]

[140] Carroll, *Alice's Adventures in Wonderland*, 4.

# Chapter Six
## Passive Attention Drawing: Belief-Driven Strategies

Standard criticisms of algorithmic curation on social media tend to offer a three-partite argumentation. Such accounts start with the assumption that recommender systems aim to present users with content that matches users' preferences. Then, the point is raised that merely tracking users' preferences traps users in an epistemic bubble: encountering *only* things one likes to see is bound to blind a person to certain information. From this, the conclusion is drawn that algorithmic curation will produce deficient belief systems in people: the bad evidential coverage which the epistemic bubble constitutes will cause certain beliefs to be disproportionately corroborated, while other beliefs are rejected on the basis of incomplete evidence.[141]

In the last two chapters I have argued that the basic assumption on which such accounts are based is incomplete: recommender systems of the kind which social media services can be expected to employ do not necessarily track users' preferences at all. In their ruthless pursuit of users' attention, these systems can also be expected to take advantage of deficiencies in the epistemic environment which social media offer in order to deceive users into believing that certain content is truthful, as in the case of fake news, or they can be expected to use intentionally misleading content as a tool to extend users' time spent on the service, as in the case of rabbit holes.

What my account has so far shared with standard criticisms of algorithmic curation is the idea that deficient belief is in the end an accidental consequence of the design of the recommender system. The pollution of people's belief systems with improperly held beliefs seems a side effect of algorithmic curation in the same way that air pollution is a side effect of heavy industry. This gives the impression that the negative effects are separable from the technology. We have a feeling that it must be possible to employ heavy industry without polluting the air, for example by "catching" the exhaust fumes before they disperse. In the same way, my account so far gives the impression that it might be

[141] This is roughly the argumentative structure of Cass R. Sunstein, *#Republic: Divided Democracy in the Age of Social Media* (Princeton ; Oxford: Princeton University Press, 2017). Sunstein places his account in a wider context of deliberative democracy, and therefore his argument goes beyond the mere worry that social media cause deficient belief systems. But the argument starts with the three-partite structure which I sketch: algorithmic social media construct a *daily me* for us; this daily me only consists of things we like; therefore on a societal scale we will see phenomena like polarisation because we do not hear the relevant counterarguments until it is too late to be persuaded by them. (See especially chapter 3, 59-97)

possible to “catch” the bad beliefs before they disperse in people’s belief systems, perhaps by training citizens in subjects like digital literacy. Essentially, this would mean that we acknowledge the opacity of social media, but that the social practice we erect around it puts the epistemic burden on the shoulders of users: it would be their responsibility not to be taken in by the misleading broadcasts which they encounter. No epistemic responsibilities would accrue to social media services, and therefore nothing would have to be changed about these services.

But there is something strange about this suggestion. An attention maximising recommender system is geared toward learning to predict the user’s behaviour in different scenarios, after which it chooses the scenario that is most likely to increase the time spent on the service. Thus, recommender systems as they are used by social media services are technologies aimed at manipulating human behaviour. But as Charles Sanders Peirce argued almost a century and a half ago, “(o)ur beliefs guide our desires and shape our actions.”[142] Regardless what we think of Peirce’s philosophy, it is hard to deny that among the most important determinants of a person’s behaviour are the beliefs which that person has. Is it then likely that a technology which is entirely geared toward the manipulation of human behaviour will do so without manipulating belief?

In this chapter, I consider whether attention-maximising recommender systems can be expected to employ the manipulation of belief as one of its tools. I do this by answering two sub-questions. First, I consider whether it is within the abilities of a recommender system to manipulate belief at all. Second, I ask whether belief manipulation can aid in achieving the specific goal for which social media recommender systems are constructed: the maximisation of human attention.

The analysis in this chapter is not meant to contradict other accounts of the harmful epistemic effects of social media services. Indeed, my analysis is perfectly compatible with most of these accounts. The added value of this chapter is rather that it proposes another way to approach this topic, that puts greater emphasis on the activities of social media services themselves. In the next and final chapter, I argue that this analysis also gives us reasons to put some epistemic responsibilities on the shoulders of these services.

[142] Charles Sanders Peirce, “The Fixation of Belief,” in *Classics of Western Philosophy*, ed. Steven M. Cahn, 8th ed. (Indianapolis/Cambridge: Hackett Publishing Company, Inc., 2012), 1249.

## 1. Can social media recommender systems manipulate belief?

Before I can start to answer this chapter's question, I must first clarify its central terminology. By "belief manipulation" I mean the activity of moulding (parts of) a person's belief system in a certain image. It is thus not quite the same as convincing someone. We generally understand the activity of attempting to convince someone as mediated by rational argumentation. Belief manipulation, by contrast, does not restrict itself to merely rational means: any available tool can be used to manipulate someone's belief system. Another contrast is that the activity of convincing someone is generally aimed at changing someone's mind, while this is only one of the aims of a belief manipulator. Thus we see that convincing is a sub-class of belief manipulation. Specifically, there are three modes of belief manipulation. Say that a would-be manipulator wishes his victim to have belief P. If the victim already has belief P, then manipulation can be performed to prevent this belief from being defeated. If the victim has the belief ~P, then manipulation can be performed to change this into P. If the victim does not have a professed belief on the topic yet, then manipulation can be performed to ensure that the victim acquires P rather than ~P.

The question whether a social media recommender system is capable of manipulating belief is best approached by breaking it into two. The first question to be answered is whether a social media recommender system has the necessary content available for manipulating a user's belief. This question might be easiest conceived of as the question whether, if content curation was performed by a human actor instead of a recommender system, this person would be able to manipulate a user's belief system. The second question to be answered is whether the strategies that this person would employ are also available to the recommender system.

Starting with the second question, remember that although recommender systems are complex software based on an array of advanced technologies, the principle behind these systems is relatively simple. A recommender system attempts to measure, and then predict, whether a given piece of content lengthens or shortens a user's time spent on the service. The content that is likely to fulfil the ODS best – in the most naïve case, the content that lengthens the time spent on the service the most – will be recommended to the user. Therefore the question whether a recommender system can fulfil its ODS by means of belief manipulation really comes down to two factors.

1. *Is the effect of belief manipulation on user attention strong enough to be clearly measured?* If manipulating a user's belief system would only have a minimal effect on the time spent on the service, then it is unlikely that the recommender system could pick up on it. In other words: it must be conclusively "worth it" to attempt to manipulate a user's belief system.
2. *What is the time scale over which increases in attention are measured?* This point is somewhat more complex, because it is obviously difficult to gauge whether a given recommendation – or a given path of recommendations – is the reason that someone spends significantly more time on the service a year later. As time progresses, more variables are introduced which might also have influenced time spent on the service. Yet being able to measure effects over a longer period of time is important, because it is unlikely that belief manipulation will happen overnight - in any case, manipulating a user's belief system will take longer than the split second needed for the decision to click this rather than that news article. Whether the recommender system will be able to untangle the effects of different variables over a longer period of time is dependent on at least two things: the size of the effect of belief manipulation, and the size of the dataset (describing the effects of past recommendations on this and other users) available to the recommender system. Given a sufficiently large effect and dataset, long-term effects will be measurable.

Both these questions are empirical, and can therefore not be answered in the present work. With regards to the second question, however, there is some indication that the designers of social media recommender systems do have time scales in mind that are longer than merely the duration of the recommended content. Specifically, YouTube employee Eric Meyerson mentions in his 2012 blog post that recommendations were geared toward videos "that increase the amount of time that the viewer will spend watching videos on YouTube, not only on the next view, but also successive views thereafter".[143] This gives a strong impression that the effect of a recommended video was monitored for longer than the duration of that single video.

It seems, then, that the answer to the question whether recommender systems have the capacities to employ strategies that are based on belief manipulation can be a

[143] Meyerson, "YouTube Now."

conditional yes. If the effects of these strategies are strong enough, if the recommender system is designed in such a way that it measures long-term effects, and if the recommender system has access to a large enough dataset to effectively do this, then these strategies are open to the recommender system. Although there is no way to tell whether these conditions are met by any contemporary recommender system, it is important to note that meeting these conditions is not contingent on technological progress. In principle, it seems that the technology for belief manipulation is already available in recommender systems as they exist today: the question is just whether the effects and available datasets are sufficiently large. This leaves unaddressed, of course, the question whether any belief-driven strategies aimed at attention maximisation actually exist: I will address that question in section three.

In the rest of this chapter, I will sometimes seem to talk about recommender systems as if they possess intentionality. My talk of strategies, for example, might give the impression that I believe recommender systems to first design a strategy and then execute it, the same way a human would. I hope that the above goes some way toward dissuading the reader from this interpretation. Water flows to the lowest point under the influence of gravity. We can predict the way water will flow by mapping the topology of its surroundings, calculating gradients, recognising obstacles. If we perform well, the water will take the exact route we predict. But it would be a mistake to say that the water took that route because it understood the gradients, because it recognised the obstacles, and knew the topology: water simply does its thing. Just like water is pulled down by gravity, so a recommender system's recommendations are shaped by attention. In order to know how the recommender system will behave, we must know the attentional context in which it operates. While mapping this attentional context I will speak in terms of strategies because that is an easy way to talk, just like talking of routes is practical when we wish to predict the way water will flow. But that does not mean that the recommender system is aware of the strategies it uses, any more than water is aware of the route it takes: the recommender system simply does its thing, and our theorising follows.

## 2. Three strategies for curation-based belief manipulation

This takes me back to my first question: are strategies for belief manipulation available? Imagine that there was a human curator responsible for curating the content available to a user through his social media feed, and that this curator had the intention to manipulate this user's belief system. How would that curator go about it? Would the curator have any epistemic weapons in his arsenal?

I think we tend to intuitively answer this question in the positive. As I mentioned in the introduction to this chapter, many of the concerns which are raised about recommender systems and social media in general are exactly dependent on the idea that these technologies can have accidental effects on people's belief systems: this also holds for my treatments of deceptive news and rabbit holes. But the description of these effects as "accidental" is an attribution of intentionality to these systems, or the designers of these systems, which is irrelevant for the question whether these systems are *capable* of having effects on belief systems: if a technology can have an accidental by-effect, then that technology can also be used to produce these effects intentionally. Any factory that pollutes the air in order to produce goods can also produce goods in order to pollute the air. Thus if we believe that recommender systems, or social media in general, can have certain bad effects on people's belief systems, then contained in that belief is the belief that the infrastructure and content of social media can be used to manipulate people's belief systems.

Nevertheless, it will be good to consider some different ways in which recommender systems can manipulate a user's belief system. Specifically, I will consider three different ways in which a recommender system can achieve this. First, there are what we might call *non-rational* strategies: strategies which are based on empirically discovered psychological effects. Second, there are *rational* strategies. Last, I will emphasise the role of human actors in this process. After all, a recommender system merely distributes content from the one to the other user. Therefore the brunt of the belief-manipulating work can be shouldered by human actors instead of the recommender system.

### *2.a) Non-rational strategies*

By non-rational strategies, I mean strategies that employ mechanisms the explanations of which do not follow the model of rational action. These are mechanisms which have been observed empirically, often in psychological research. Psychological literature is rife with such mechanisms, and this is not the place to list all of them. I will focus instead on three well-known mechanisms and show how these mechanisms can be employed by an epistemic manipulator. It goes without saying that psychological science is not finished, and there are good reasons to believe that more of such mechanisms may exist which are yet undiscovered. The discussion of these three mechanisms serve as an example of how, in general, a recommender system might take advantage of the bundle of heuristics that go hand in hand with human cognition.

The first candidate is the illusory-truth effect: encountering repeated utterances of the same proposition increases one's confidence in the truth of that proposition. This effect, studied both within psychology and consumer research,[144] is robust across different types of situations. One of the few constraints on the effect is that one needs to be unsure about the truth of the proposition to begin with: if one's mind is already clearly made up, then mere repetition will not do much to change it.[145] Interestingly, the effect even occurs when the statement derives from a source that is known to be untrustworthy.[146]

The illusory-truth effect is part of a larger class of fluency effects.[147] Processing fluency is defined as "the subjective experience of ease with which people process information".[148] This experience of ease has been shown to be used as a heuristic for judging the validity or truth of a claim: the easier it is to process a claim, the more likely that claim will be judged "true".[149] This heuristic can be problematic because it is possible to increase processing fluency in a variety of ways that have no bearing on the truthfulness of a claim: Nguyen's concerns about the seductions of clarity were partially based on this insight.[150] Fluency effects have been observed in a variety of experiments where processing fluency was increased by many different means.[151] In the previous paragraph I mentioned repetition as one of these means, but processing can also be eased by using rhyming statements,[152] or by including a picture along with the claim.[153] Indeed,

[144] Alice Dechêne et al., "The Truth About the Truth: A Meta-Analytic Review of the Truth Effect," *Personality and Social Psychology Review* 14, no. 2 (May 2010): 238, https://doi.org/10.1177/1088868309352251.

[145] Dechêne et al., 239.

[146] Ian Maynard Begg, Ann Anas, and Suzanne Farinacci, "Dissociation of Processes in Belief: Source Recollection, Statement Familiarity, and the Illusion of Truth.," *Journal of Experimental Psychology: General* 121, no. 4 (1992): 446.

[147] Dechêne et al., "The Truth About the Truth," 238.

[148] Adam L. Alter and Daniel M. Oppenheimer, "Uniting the Tribes of Fluency to Form a Metacognitive Nation," *Personality and Social Psychology Review* 13, no. 3 (August 2009): 219, https://doi.org/10.1177/1088868309341564.

[149] Alter and Oppenheimer, 219.

[150] Nguyen, "The Seductions of Clarity," 237.

[151] Alter and Oppenheimer, "Uniting the Tribes of Fluency to Form a Metacognitive Nation," 220.

[152] Matthew S. McGlone and Jessica Tofighbakhsh, "Birds of a Feather Flock Conjointly (?): Rhyme as Reason in Aphorisms," *Psychological Science* 11, no. 5 (September 1, 2000): 426–27, https://doi.org/10.1111/1467-9280.00282.

[153] Eryn. J. Newman et al., "Truthiness, the Illusory Truth Effect, and the Role of Need for Cognition," *Consciousness and Cognition* 78 (February 2020): 2–3, https://doi.org/10.1016/j.concog.2019.102866.

in a comprehensive review of fluency effect literature, Adam L. Alter and Daniel M. Oppenheimer argue that "fluency exerts the same influence on judgments independently of how it is generated".[154]

Lastly, there are indications that believing that people with whom we identify hold a certain view can have a powerful influence on our willingness to accept that view.[155] For example, when judging the merit of certain policies, American study participants overwhelmingly sided with the policies they were told had been proposed by the political party they sided with – almost entirely disregarding the content of the policies. Judgements were markedly different if participants were not told which party had proposed the policies.[156] Some disagreement exists over whether this mechanism is rational or not. Rebecca Rini, for example, argues that in some contexts it is reasonable to defer to the judgement of people who share the same partisan affiliation, because similar partisan affiliation can be an indicator of sharing values which are relevant to the matter at hand.[157]

I mention the illusory-truth and fluency effects because they are easily employed by a curator both in the sense that it is likely that the necessary content is available and in the sense that it is simple to induce the effects. For the illusory-truth effect, the only thing required in terms of content is that there are multiple instances of content that make the same or similar claims. For other fluency effects, what is required in this respect is simply easy-to-process content – for example content accompanied by a picture (think of memes), or content that expresses its point in a very simple manner.

I mention the identity effect because this effect is particularly likely to have powerful applications in the mixed type of social media services that we see today, where curation is performed both by means of social network functionality and by means of recommender systems – see chapter one for a more elaborate discussion of this type of social media services. On Facebook, for example, it is not uncommon to be presented with a piece of content along with the message that a certain social connection interacted with it. Given the potential strength of the identity effect, this can function as a powerful

[154] Alter and Oppenheimer, "Uniting the Tribes of Fluency to Form a Metacognitive Nation," 220.
[155] Neil Levy, "The Bad News about Fake News," *Social Epistemology Review and Reply Collective* 6, no. 8 (2017): 25.
[156] Geoffrey L. Cohen, "Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs.," *Journal of Personality and Social Psychology* 85, no. 5 (November 2003): 819, https://doi.org/10.1037/0022-3514.85.5.808.
[157] Rini, "Fake News and Partisan Epistemology," E49–54.

strategy for belief manipulation, for example by showing many such reports of friends liking content that supports P while withholding such reports of friends liking content that supports ~P.

These effects become even more powerful when they are combined. For example, there is evidence that fluency effects play a role when we attempt to estimate the popularity of a certain view in a group of peers. Specifically, if a person was exposed to repeated utterances of the same claim *by the same member of a certain group*, then that person's estimation of the popularity of that view in the entire group increased.[158] This finding serves to link the fluency effect to the identity effect: not only can we expect repeated exposure to the same claim to increase a user's estimation of the truth of that claim directly, it might also do this by increasing the user's estimation of the popularity of that claim in the group he identifies with, thereby triggering the identity effect which will also add to the user's confidence in said claim.

This short discussion has shown how easy it is for a curator to take advantage of human cognitive heuristics in order to manipulate a user's belief system in this or that direction. Importantly, a recommender system's ability to take advantage of a certain non-rational mechanism is independent of human knowledge of that mechanism: these strategies are not consciously programmed into the system, but rather discovered by the system itself by means of trial and error. This discussion is therefore necessarily limited, because in the end there is no way of knowing *why* a recommender system recommends a certain piece of content: the system simply does what has worked in the past, without explaining to itself why that happened to work.

### *2.b) Rational strategies*

Explanations of the type discussed in the previous subsection are popular in discussions of the dangers of social media and other types of services driven by artificial intelligence.[159] In many of these discussions, the human brain is represented as a manipulable, evolved contingency that simply does not stand a chance against the

[158] Kimberlee Weaver et al., "Inferring the Popularity of an Opinion from Its Familiarity: A Repetitive Voice Can Sound like a Chorus.," *Journal of Personality and Social Psychology* 92, no. 5 (May 2007): 831, https://doi.org/10.1037/0022-3514.92.5.821.

[159] For example, this is the perspective that is generally taken by the Center for Humane Technology. See Center for Humane Technology, "How Social Media Hacks Our Brains"; or the various utterances made in the Center's film: Orlowski, *The Social Dilemma*; or its podcast: Harris and Raskin, "Your Undivided Attention."

‘supercomputers’ that social media services use. Recommender systems, according to this account, exploit pre-existing weaknesses of the human cognitive apparatus. I believe this to be a limited and probably incorrect way of explaining what is going on, for two reasons. First of all, there is no consensus that non-rational mechanisms of the kind which I just discussed should be considered weaknesses at all – or even irrational, for that matter. It is clear that these mechanisms ought not be used by an epistemically perfect agent that has infinite resources, but it is also clear that we are not such an agent. Given the fact that humans are limited creatures, the rational pathways available to us must also be limited. Therefore it might be perfectly rational *given the limited circumstances* for humans to use exactly the kinds of heuristics represented by the mechanisms which I just discussed, as long as these mechanisms are broadly reliable. [160]

Secondly, and more importantly, even perfectly rational thought does not always lead to truth. Remember that Goldberg separated epistemic propriety from knowledge: even beliefs that are held responsibly are not guaranteed to be correct.[161] Given particularly bad epistemic circumstances – for example, untrustworthy peers who provide one with bad evidence – a community of rational actors might be misled. This is nicely illustrated by formal modelling in social epistemology, sometimes called the network epistemology approach.[162] In these types of studies, communities of knowers are represented by nodes in a network. Every node is faced with a dilemma between options, A and B, which it needs to resolve based on the evidence which is available to it. In the first accounts of this type, introduced by Kevin Zollman,[163] the network represents a community of researchers. Every researcher has to decide which option to study – A or B – and makes this decision by using Bayes’ rules to judge which option is most promising given the available evidence. By studying an option, the researcher produces more evidence, which it can then share with some or all of the other researchers. Because not all study results necessarily point in the direction of truth due to normal statistical variations, misleading evidence can come to the fore which points in the wrong direction. Under certain

[160] For two recent discussions of this idea of “bounded rationality”, see Morton, “Human Bounds”; and Dallmann, “When Obstinacy Is a Better (Cognitive) Policy.” See also Gigerenzer and Gaissmaier, “Heuristic Decision Making.” for a broader overview of the reliability of these kinds of heuristics.

[161] See chapter three, section four of this work.

[162] Cailin O’Connor and James Owen Weatherall, “Modeling How False Beliefs Spread,” in *The Routledge Handbook of Political Epistemology*, ed. Michael Hannon and Jeroen de Ridder, Routledge Handbooks in Philosophy (Abingdon, Oxon ; New York, NY: Routledge, 2021), 205.

[163] Kevin J. S. Zollman, “The Communication Structure of Epistemic Communities,” *Philosophy of Science* 74, no. 5 (December 2007): 574–87, https://doi.org/10.1086/525605.

circumstances, such misleading evidence can cause a cascade effect which causes the whole research community to switch to the wrong option. At this point, no more evidence is produced regarding the secretly better option, and therefore the community is effectively trapped in its decision.

James Owen Weatherall, Cailin O'Connor, and Justin P. Bruner introduced an interesting variation on this theme.[164] In their model, the network has three different kinds of nodes: there is a community of researchers as was described above, a community of policymakers, and a propagandist. The policymakers can be understood to represent government officials, or even democratic citizens: their assignment is to establish the better option, just like the researchers, but they do not have the capacity to produce new evidence of their own. Instead, they receive evidence from some or all of the researchers. The propagandist is a single node which observes the activities of all the researchers, and communicates with all the policymakers. Its goal is to cause the policymakers to choose what is actually the worse option – call it B. It promotes B by selectively communicating results to the policymakers: of course, the propagandist will only communicate results that support B over A.[165] The introduction of this single propagandist has enormous effects on the way the model develops. In many cases, the propagandist is utterly successful in its goal: "while the community of scientists converges on true beliefs about the world, the policymakers reach near certainty in the false belief."[166] This is all the more startling because the epistemic circumstances are really not that vicious: all policymakers use apparently rational Bayesian rules for their judgements, and all the evidence they base their judgements on is the result of proper scientific practice. Nevertheless, the policymakers are, in many cases, misled.

The propagandist reaches its goal by shifting the weight of the evidence in favour of a certain conclusion. Important in this respect is that the propagandist is a mere addition to a policymaker's epistemic network: the presence of the propagandist does not change the fact that the policymakers are still connected to real, trustworthy scientists. The propagandist achieves its end by supplying the policymakers with more – albeit skewed – evidence.

[164] James Owen Weatherall, Cailin O'Connor, and Justin P. Bruner, "How to Beat Science and Influence People: Policymakers and Propaganda in Epistemic Networks," *The British Journal for the Philosophy of Science* 71, no. 4 (December 1, 2020): 1157–86, https://doi.org/10.1093/bjps/axy062.
[165] Weatherall, O'Connor, and Bruner, 1162.
[166] Weatherall, O'Connor, and Bruner, 1164.

Compare the role of the propagandist to the role of a social media curator. Surely, the curator exerts at least as much influence over the evidence available to a certain user as the propagandist does over the evidence available to a certain policymaker. Indeed, the curator's influence is arguably stronger: unlike the propagandist, the curator can decide that certain evidence will not reach the user at all – at least through the social media service. If the propagandist can lead a community of rational policymakers entirely astray by means of its limited abilities, then it seems likely that a social media curator can achieve the same.

### *2.c) The role of human actors*

One aspect of belief manipulation on social media has still been underemphasised: the role of human actors.

The role of the social media curator, although important, is relatively limited. The curator does not create content; it merely decides how it is to be distributed. The art of changing people's minds is an old one, however, and it is well understood. Many books on the topic have been published, starting with or even before Aristotle's *rhetoric*,[167] and all this information is available to those who produce content: social media users. These users are not unbiased researchers, merely motivated by the honourable intention to spread truth among their peers. There are many reasons why we might wish to convince others of the truth of a certain proposition, ranging from the political to the financial, from the well-intentioned to the vicious. The sheer number of users that the more popular social media services enjoy makes it likely that all these various motivations to convince have been translated into a practically infinite collection of misleading content, which employs all the tricks known to human kind in order to convince its consumers of whatever its originator wished to convince people of. Therefore the social media curator has easy pickings: it does not need to employ elaborate strategies in order to manipulate its users' beliefs systems – it merely needs to identify content which employs these strategies.

This latter fact will be extra salient in cases where the interests of content creators and social media services align: if content creators have an incentive to produce content that manipulates its consumers' belief system in such a way that it maximises attention,

[167] Aristotle, *Rhetoric* (South Bend, United States: Infomotions, Inc., 2000), http://ebookcentral.proquest.com/lib/uunl/detail.action?docID=3314386.

then the social media curator will have access to a great arsenal of content that helps it achieve its goal, as well. Of course, social media services are often set up in such a way as to achieve exactly this alignment of interests. For example, both Facebook and Youtube pay video creators to include ads in their videos, which generate more revenue when more people view them.[168] Social media services intentionally offer an infrastructure that ensures that not only they, but also all their users work in the context of the attention economy. In section 3.d I will briefly revisit the role human actors can play in attention-maximisation by means of belief manipulation.

## 3. Attention-maximising belief systems

Multiple strategies for belief manipulation are open to social media curators, and these strategies are available to recommender systems provided that – among other conditions – they have a sufficiently strong effect. In the case of attention-maximising recommender systems, this means that the belief systems which are propagated by means of these strategies must significantly increase the amount of attention spent on the service. This raises the question: what kind of belief system can have such an effect?

### *3.a) Trust as the object of manipulation*

The most obvious candidate would be an explicit belief to the effect that the recommender system's ODS must be fulfilled. This would amount to something like "It is good to spend a lot of time on social media." Such a belief seems an obvious candidate not only because it effectively achieves its goal, but also because it is simple: successful implementation is merely dependent on the implantation of a single belief. But closer consideration reveals that this simplicity, far from being an advantage, is actually a great hindrance to implementation. Beliefs never come alone, but are rather supported by each other: I believe that I will get wet if I go outside *because* I see that the streets are wet, and *because* I know that the last time it rained, the streets were wet too, and *because* I remember that the last time I went outside when it rained, I got wet as well. Attempting to implant a single belief is like trying to only build the spire of a church: if it has nothing to rest on, it is unclear how to even begin. The belief that it is good to spend a lot of time

[168] "How to Earn Money on YouTube - YouTube Help," accessed June 12, 2021, https://support.google.com/youtube/answer/72857?hl=en; "How to Make Money from Your Content on Facebook," Facebook for Business, accessed June 12, 2021, https://en-gb.facebook.com/business/learn/lessons/how-make-money-facebook.

on social media is so specific that it needs a very particular foundation. There might be a small subsection of the user population that has a belief system that is hospitable to this belief, but most users' belief systems will need to be radically altered before they are ready to accept it.

To ensure ease of implementation across different types of users with radically different belief systems, it might be better to consider, instead of particular beliefs, structural features that a belief system can have which would increase the attention spent on the service. If such structural features exist, they might be filled in with the beliefs that fit a particular person best: if you are a user of type α, then your belief system will be manipulated to approximate model A; if you are a user of type β, then your belief system will be manipulated to approximate model B. The beliefs implied by models A and B can be radically different, but the models share certain structural features that maximise attention.

It is widely accepted that our belief systems must indeed include certain features that regulate our attention, although it is not common to describe them as such. In the literature, these features are more commonly described by means of the term "(epistemic) trust".[169] Epistemic trust is best understood, argues Katherine Dormandy, as a three-place relationship between a hearer, a source and an object: I trust you to supply me with information about a certain topic.[170] The topical restriction of the object is common, but not necessary: the average person might trust a baker to give her accurate information about bread, but not about the metaphysical nature of the universe; a member of a deeply religious community, however, might trust that community's priest on both these topics.

Epistemic trust does not directly influence attention. It is perfectly possible to pay attention to someone we do not trust. Yet when we stop trusting the one and start trusting the other person, it is likely that the focus of our attention will change as well. We might still pay some attention to the person we now deem untrustworthy, but we have much less reason to do so: if we are motivated to learn the truth about the world, then we would do better to direct the majority of our attention at the more trustworthy person.

[169] For some discussions of epistemic trust, see Katherine Dormandy, *Trust in Epistemology* (New York; London: Routledge, 2020); Baurmann, "Fundamentalism and Epistemic Authority."

[170] Katherine Dormandy, "Introduction: An Overview of Trust and Some Key Epistemological Applications," in *Trust in Epistemology*, ed. Katherine Dormandy (New York; London: Routledge, 2020), 2.

The relationship between epistemic trust and attention is one reason why the manipulation of epistemic trust would be an interesting strategy for an attention-maximising belief manipulator. A second reason is what we might call the transitivity of epistemic trust. Michael Baurmann, whose work I briefly mentioned in chapter four, argues that there are two types of epistemic trust: social trust, which is based on general social conventions, and personal trust, which is based on our own judgements of the trustworthiness of other people. Such judgements can be based on all kinds of evidence, ranging from certain personality traits of the would-be trustee to evidence of his or her actual epistemic achievements.[171] Personal trust judgements tend to trump social trust.[172] Especially interesting at present is that personal trust judgements do not only cause me to strongly trust my specific trustee: "I will also be inclined to ascribe a comparable high trust-value to information which stems from sources whose trustworthiness has not been ascertained by myself, but by the testimony of people I personally trust."[173] This transitivity of personal trust is part of the reason why personal trust can be so powerful: a single trust-judgement can open me up to a vast range of new information. But the very strength of this property also makes it an extremely useful backdoor for would-be epistemic manipulators: if a victim can be manipulated into trusting a single well-placed person, this can open up the victim's belief system to a vast network of people who espouse exactly the kind of beliefs the manipulator wishes his victim to acquire. Thus a tiny, well-placed manipulation of a single element of the victim's personal trust-network can have far-reaching consequences for his entire belief system.

In chapter five, I introduced the story of Caleb Cain – the college dropout who, in his own words, fell down the alt-right pipeline. Interestingly, Caleb's descent into the land of the alt-right followed exactly this pattern. His pursuits on YouTube started innocently: he watched self-help videos to help lift him out of his depression. But his favourite YouTubers would introduce him to new ones that were more extreme – for example by interviewing them on their shows – and step by step the videos he watched became more radical. Eventually, he found himself watching material that he would never have considered watching when he started. Caleb was lured in by the transitivity of trust. In

[171] Baurmann, "Fundamentalism and Epistemic Authority," 60.
[172] Baurmann, 61.
[173] Baurmann, 61.

his own words: "You take one piece of it, and that will take you to the next step, and the next step. And this is how you become radicalised."[174]

### *3.b) Echo chambers as attention monopsonies*

In their book *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*, politics and media scholars Kathleen Hall Jamieson and Joseph N. Capella describe the structure of the then still young conservative media establishment in the United States.[175] They describe this structure as an "echo chamber": different parts of the establishment echo each other's claims, providing mutual support for each other's utterances and legitimising each other as trustworthy sources.[176] Not only do they provide their audiences with positive reasons to trust other parts of the echo chamber: they also undermine the trustworthiness of sources outside of the conservative media establishment. They achieve this latter feat by espousing a narrative that the mainstream media are liberal, that liberals are untrustworthy, and that this is shown by the fact that these liberal media employ a double standard: they hold conservatives accountable for offences that they let slide whenever a liberal commits them.[177] Because the mainstream media are untrustworthy, truth-loving people would do better to only trust the conservative media establishment. Together, the features of the echo chamber have the effect of insulating the conservative audience by means of a manipulation of epistemic trust.

Jamieson and Capella argue that the echo chamber structure of the conservative media establishment is a fruitful political strategy: it aids the Republican party in keeping together the coalition of voters that brought Reagan to power.[178] It is likely, however, that the construction of the conservative echo chamber was motivated by economic incentives as much as it was by political ones. Two of the three programs that Jamieson and Capella analyse for their book are owned by Rupert Murdoch, a man who "has built a media empire on the realization that there is commercial value in creating media outlets that tilt to the right."[179] The creation of an echo chamber can be seen as an extreme form of what is known as "product differentiation" in economic science, a process where social norms

[174] Cain, *My Descent into the Alt-Right Pipeline*.
[175] Jamieson and Cappella, *Echo Chamber*.
[176] Jamieson and Cappella, 76.
[177] Jamieson and Cappella, 37–38.
[178] Jamieson and Cappella, 56–74.
[179] Jamieson and Cappella, 43.

are manipulated in order to make close-to-identical products sufficiently different in the eyes of the consumer for a single producer to monopolise a part of the market. Luxury clothes brands are a good example: although their products are virtually identical to those of most other brands, luxury brands have managed to imbibe their brands with such an air of luxury and exclusivity that they can charge significantly higher prices than their competitors. Similarly, the echo chamber of the conservative media establishment creates an intense kind of 'brand loyalty' that ties viewers to these outlets.

When we see things from an attention-economic perspective, however, we need to make a slight change to this account. In the case of product differentiation, the supplier differentiates the product in order to increase demand by creating an effective monopoly. But from an attention-economic perspective, the viewer is the *supplier* of attention while the demand side is represented by the media establishment. Thus an echo chamber does not establish a monopoly – which is a market structure with a single supplier – but rather a monopsony: a market structure with a single consumer. Seen from the attention-economic perspective, echo chambers are closer to oil rigs than to luxury brands: they serve to keep other would-be extractors of attention out of the picture.

### *3.c) Echo chambers on social media*

I have argued that the construction of an echo chamber was a successful strategy for the conservative media not only from a political, but also from an economic perspective. Could the same formula be successfully implemented on social media? And under which conditions?

We first need a generalised definition of echo chambers that goes beyond the case of the conservative media establishment. Thi C. Nguyen provides such a definition. It is based both on Jamieson and Cappella's account and on insights from social epistemology, and is therefore perfectly tailored to the present inquiry. Nguyen writes:

> *I use "echo chamber" to mean an epistemic community which creates a significant disparity in trust between members and non-members. This disparity is created by excluding non-members through epistemic discrediting, while simultaneously amplifying members' epistemic credentials. Finally, echo chambers are such that*

> *general agreement with some core set of beliefs is a prerequisite for membership, where those core beliefs include beliefs that support that disparity in trust.*[180]

On Nguyen's account, echo chambers have two defining elements: there is a disparity in trust between members and non-members, and membership is defined by agreement with some core set of beliefs. The conservative media establishment's strategy was so fruitful politically because it hooked on to a pre-existing core set of beliefs: conservative politics existed before the rise of the conservative media. By positioning themselves in such a way that the line dividing trustworthy from untrustworthy which was implied by this core set of beliefs coincided with the line dividing the conservative news establishment from its competitors, they managed to turn this political strategy into an economical one.

Whereas the conservative media establishment espouses a single and coherent narrative, this would be impossible for a social media service to achieve. This is because the design of social media services separates the role of content creator from the role of distributor. An echo chamber that benefits a social media service can therefore not make the same contrast between trustworthy and untrustworthy media without thereby also excluding a portion of the social media service itself. While the dividing lines between liberal and conservative coincided with the lines separating CNN from Fox, that same division between conservative and liberal cuts right across Facebook. But this is no problem for Facebook, for the goal of attention maximisation is not to increase a single user's attention on all parts of Facebook, but rather to increase that user's attention on Facebook in general: it does not matter whether a specific user only listens to right-wing or left-wing sources, as long as all the sources that he listens to espouse the majority of their views through Facebook.

Thus there is really only one requirement for an echo chamber to be beneficial from the perspective of a social media service, and that is that if the dividing line between trustworthy and untrustworthy does not cut across the service, it cuts between the service and other media. An echo chambers aids in maximising the attention that is directed at a social media service if it actively discredits sources that espouse their views in other places than the social media platform.

[180] C. Thi Nguyen, "Echo Chambers and Epistemic Bubbles," *Episteme* 17, no. 2 (2020): 146, https://doi.org/10.1017/epi.2018.32. I omitted Nguyen's italicised emphases to improve legibility.

A natural objection to this argument arises from Nguyen's distinction between echo chambers and epistemic bubbles. According to Nguyen, "(a)n epistemic bubble is a social epistemic structure which has inadequate coverage through a process of exclusion by omission."[181] Epistemic bubbles blot out parts of the available information. This is clearly different from an echo chamber: a member of an echo chamber might still get all the same information that a non-member gets, but she heavily distrusts all information that does not originate from within the chamber. But if this is the case, then would it not be a better strategy, from an attention perspective, to construct an epistemic bubble than to construct an echo chamber?

I think this objection misguided because it rests on the mistaken assumption that echo chambers and epistemic bubbles are mutually exclusive in some way. But they are not: the two concepts involve descriptions on two different levels. While the concept of epistemic bubbles describes which information actually reaches a certain person, the concept of echo chambers describes the mechanisms by means of which trust is distributed by a group of people. Thus it is perfectly possible that someone finds herself in an epistemic bubble *because* she is part of an echo chamber: she might believe the central tenets of a conspiratorial group and therefore only consume information which is distributed by members of that group. Merely saying that someone finds herself in an epistemic bubble is not that informative: it just means that she finds herself in a social epistemic structure with spotty coverage. Thus, I argue, it is not clear what "constructing an epistemic bubble" means: the term only serves to describe the effect of an attention-maximising strategy, but not the strategy itself. It can therefore not serve as an *alternative* to constructing an echo chamber, which *is* a description of a strategy.

### *3.d) Echo chambers and human actors*

In section 2.c I argued that recommender systems can manipulate beliefs by enlisting the activities of human belief manipulators. Such enlisting would be especially likely, I argued, if the incentives of social media services and users aligned: if users stand to gain from maximising the attention they receive from other users, then they can be expected to produce exactly the kind of content the recommender system needs in order to manipulate user beliefs in its favour.

[181] Nguyen, 143.

Now that my account has zeroed in on the construction of echo chambers as a fruitful attention maximising strategy, I wish to briefly revisit this point, for the construction of echo chambers is, I think, a relatively common activity which human actors engage in for a variety of reasons. We saw before that the conservative media establishment constructed an echo chamber for political and economic reasons. Constructing an echo chamber is also a staple of cult indoctrination,[182] as well as being a core part of many conspiracy theories.[183] And certainly this last kind of echo chamber can be extremely lucrative: American conspiracy theory peddler Alex Jones, for example, has constructed a successful business empire by selling a variety of products especially targeted at those who hold the core beliefs of his echo chamber to be true.[184]

This is all to bring out the point that it ought not be too difficult for a social media recommender system to maximise attention by means of echo chambers. These systems do not need to engage in the difficult business of actually constructing echo chambers themselves: (attempts at) echo chambers are so ubiquitous in human culture already that all the recommender system needs to do is promote them.

### Conclusion

I started this chapter by asking whether belief manipulation, rather than being merely an accidental by-effect of algorithmic curation, might be one of the tools which a recommender system has at its disposal in its quest to maximise user attention. In the first section, I established some conditions which have to be met in order for a recommender system to be able to use belief-driven strategies. Because a recommender system does not have our capacity to conceptually map the terrain in which it operates, these conditions boiled down to the requirement that the effect of belief manipulation is measurable for the recommender system. This means that both the effect and the system's capacities need to be sufficiently strong.

[182] Nguyen, 142.

[183] Nguyen, 148.

[184] H Van den Bulck and A Hyzen, "Of Lizards and Ideological Entrepreneurs: Alex Jones and Infowars in the Relationship between Populist Nationalism and the Post-Global Media Ecology," *International Communication Gazette* 82, no. 1 (February 2020): 51–52, https://doi.org/10.1177/1748048519880726; Elizabeth Williamson and Emily Steel, "Conspiracy Theories Made Alex Jones Very Rich. They May Bring Him Down.," *The New York Times*, September 7, 2018, sec. U.S., https://www.nytimes.com/2018/09/07/us/politics/alex-jones-business-infowars-conspiracy.html.

In the second section, I ensued to formulate some strategies that a recommender system might use to manipulate belief. These divided into three not mutually exclusive categories: non-rational strategies, rational strategies, and strategies that take advantage of the persuasive capacities of content-producing users.

Having thus established that there seem to be ways for a recommender system to manipulate its users' belief systems – provided that certain conditions are met – in section three I considered what kinds of belief systems maximise user attention. I argued that we are not looking for a specific belief to be implanted, but rather a certain structure. The relevant structural features, I argued, are likely those related to trust, for our beliefs about trust regulate to a great extent the way we direct our attention. An ideal attention maximising belief structure is therefore an echo chamber: a structure that radically restricts the notion of trustworthy source to a small group of people that espouse similar views.

The echo chamber structure is such a practical strategy for attention maximisation because it easily parasitises on our pre-existing beliefs. This means that the area of application for this strategy is not limited to people who hold certain views: although the paradigm example of an echo chamber is the conservative media establishment, a similar structure might be erected around liberal beliefs. Thus, given a user with a certain set of beliefs, there will always be an echo chamber that is "epistemically close" to that user: a conservative might be easily manipulated into a conservative echo chamber, while a liberal will be herded into the liberal variation.

The dystopian picture that arises from this analysis is of a social media service fragmented into echo chambers. Members of a given chamber do not talk to members of the others, and distrust is rife. New users of the service are quickly analysed, categorised, then herded into the closest echo chamber in order to better extract their attention. Although it will offer little solace, there remains one thing which all users can mutually agree on: the mainstream media are not to be trusted.

Is this picture merely a fantasy, a sketch of our future, or even our current reality? Lacking public access to the source codes of social media services' recommender systems, there is no way to tell. If the picture is a fantasy, however, it cannot be a distant one: if my argumentation is correct, then we are at least well on our way toward possessing the necessary technology to make it into a reality. An indeed, there are reasons to think that we are at least on our way toward making the picture a reality. In an age of rising polarisation, war cries against the trustworthiness of mainstream media sound eerily familiar. Moreover, a recent study of the echo chamber effect on social media found that "platforms organized around social networks and news feed algorithms, such as Facebook and Twitter, favor the emergence of echo chambers."[185] All in all, the dystopia sketched in this chapter might be closer than we would like to think.

[185] Matteo Cinelli et al., "The Echo Chamber Effect on Social Media," *Proceedings of the National Academy of Sciences* 118, no. 9 (March 2, 2021): 2, https://doi.org/10.1073/pnas.2023301118. An important sidenote to this point is that this study used a slightly different definition of echo chambers, defining an echo chamber in terms of homophily – the degree to which members of the echo chamber have the same leanings and opinions. I believe this definition to be sufficiently close to Nguyen's definition to consider the two interdependent: groups with a high degree of homophily can be expected to have among their views certain views that go some way towards erecting a trust-based echo chamber. Further research on this matter is certainly required, however.

# Chapter Seven
# The Epistemic Responsibilities of Social Media Services

In chapter three I constructed an account of the relationship between media and epistemic responsibilities. In chapters four through six I descended from that highly theoretical vantage point to the messy real world, considering the epistemic effects that social media services might have on their users. Now it is time to ascend to more theoretical spheres again in order to answer the central question of this work: what are the epistemic responsibilities of social media services?

This chapter mostly serves to summarise points that were made over the course of this work, and to draw conclusions from them. The argument therefore proceeds at a brusque pace: I refer to the relevant chapters for more in-depth treatments of the different components of which this chapter consists. I start by revisiting my notion of media responsibility and transparency. Then I feed the conclusions from chapters four, five, and six into these notions. My central theses are twofold: first, social media are opaque, for which reason a social-epistemic practice must be erected that distributes epistemic responsibilities over social media services and those who engage with them; and if social media wish to keep fulfilling the societal role which they currently occupy, then some epistemic responsibilities must accrue to the services themselves.

## 1. The opacity of social media

My account of the relationship between media broadcasts and epistemic responsibility is based on the insight that media broadcasts can affect an audience's ability to meet the demands of epistemic propriety by changing the context in which an audience perceives a certain event: the curatorial power which media wield can be characterised as epistemic. This insight led me to distinguish transparent media from opaque media. The difference between the two is that media of the former category wear on their sleeves the way they distort their audience's perceptions. This openness allows audiences to adjust their belief-forming processes accordingly, thereby retaining the reliability of their cognitive processes. Opaque media, on the other hand, are not open about their distortions in this way, and therefore they provide no guarantee that their audience's cognitive processes remain reliable when they are based on the medium's broadcasts. In order to retain reliability, a social-epistemic practice must be erected in this case that distributes certain epistemic responsibilities over both the service and the audience.

I identified an intuitive argument why we might expect social media services to be transparent, which I called the argument-from-preference-tracking. Social media recommender systems, the argument goes, wield their curatorial power only by tracking user preferences. But this means that social media services merely ease a process that would have happened anyway: since the system tracks a user's preferences, it should only recommend content which the user is already looking for. If this is the case, then we might expect the service not to affect the reliability of users' cognitive processes, because the service does not change which information reaches the user.

I argued that this argument is mistaken, because recommender systems are not made to track preferences: social media recommender systems are constructed to maximise attention, and this goal can be achieved in a variety of ways. In chapter four I showed that recommender systems can sometimes recommend content that flies in the face of user preferences, and in chapter five I showed that recommender systems sometimes maximise attention by tracking content about which we are curious, rather than content for which we have a preference.

I also attacked the conclusion of the argument – that social media are transparent – in multiple ways. In chapter four I argued that the epistemic environment which social media offer hinders users' ability to correctly identify products of the social epistemic practice called "journalism". This hindered ability starts a dynamic process where the poor epistemic environment and the activities of the recommender system incentivise the production of more misleading news, thereby rendering the social media service increasingly unfit for news consumption. This constitutes one way in which social media services render unreliable certain cognitive processes based on the service's broadcast.

A second way in which algorithmic curation renders unreliable users' cognitive processes was explored in chapter five. In that chapter I argued that at times when the goal to maximise attention is at odds with the epistemic goal of ensuring the reliability of users' cognitive processes, the recommender system can be expected to choose for the former at the expense of the latter. In order to sustain a user's fall down a rabbit hole, the recommender system will sometimes recommend certain content *because* it is misleading. Important with respect to considerations of transparency is the fact that this strategy only works if the user is unaware that the system is employing the strategy: if the user is aware that a piece of content is recommended because of its misleading qualities, then the strategy will not work, because the user will not trust the content. This is therefore a perfect example of the opacity of social media services: that the service can function in the way it does partially depends on the service not being open about the way it might be misleading. The functioning of the service depends on the service's opacity.

All the discussions in chapters four, five, and six are speculative to some degree, because – as I mentioned in chapter two – social media services keep secret how exactly their recommender systems function. But rather than weakening it, I believe this serves to strengthen my account. My analyses in the second part of this thesis show that recommender systems, if they are constructed in a certain way, can be expected to have certain epistemic effects. The fact that we do not know exactly how contemporary social media services are constructed makes it impossible to predict with certainty the epistemic effects which these systems might have, which in turn makes it impossible to adjust our cognitive processes to these effects. Thus the secrecy around social media recommender systems makes social media services even more opaque.

## 2. Distributing responsibilities

Given the opacity of social media, a social-epistemic practice needs to be erected that places epistemic expectations (and therefore responsibilities) on both users and service. The goal of this practice is to salvage the ability of those who interact with social media to hold beliefs while meeting the demands of epistemic propriety. This leaves open the question who should shoulder the brunt of these responsibilities. One solution, for example, could be to erect a practice that puts a large responsibility on the shoulders of users, for example by training citizens in subjects like digital literacy.

But in the absence of substantial epistemic responsibilities on the shoulders of social media services themselves, we should expect the lessons of these trainings in digital literacy to be harsh. In order to solve the problems which I identified on the news market, the best solution would be to teach people only to trust articles from outlets whose reputation is familiar. But this would significantly restrict the added value of using a social media service when it comes to news: given this lesson, it might be better to just check the websites of those newspapers whose reputation you know. Solving the problem with rabbit holes requires an even more substantial intervention. Since it is an essential part of the YouTube rabbit hole that the user does not know when she is being misled, there is no guarding against this by being extra vigilant. Indeed, the system reacts when user behaviour changes: if the user becomes more vigilant, we can expect the system to recommend content that misleads in a more subtle way. The conclusion, I think, must be that the user should be warned against forming any beliefs at all whilst falling down a rabbit hole. But since a fall down a rabbit hole is sustained by *curiosity*, which essentially includes a desire to *know*, this is an impossibility result: if one is disallowed from forming beliefs, then one cannot achieve knowledge. Thus the lesson becomes: do not fall down rabbit holes on social media.

Even more detrimentally, chapter six shows that even if these lessons were perfectly internalised by all users, it would not be enough to salvage epistemic propriety. The most important lesson from chapter six is that it is possible that a recommender system has an epistemic programme of its own: the system might have the tools to manipulate users' beliefs, and stands to gain from doing this. This means that *any* recommendation that one receives from the recommender system might be in service of this epistemic programme: it might aim to manipulate a user's beliefs in a certain direction. Given that a user cannot know when this is the case because a recommender system's 'motives' are inaccessible, the lesson that must be drawn from this is: *do not form beliefs based on what you see on social media, for you cannot know whether it aims to manipulate your belief system in an inappropriate way.*

One objection might be raised here: even if we grant that recommender systems might aim to herd their users into echo chambers, why is this epistemically wrong? Jennifer Lackey, for example, argues that there is nothing intrinsically problematic about echo chambers: she argues that echo chambers are only bad if they are unreliable, in the sense of causing many false beliefs.[186] So unless we can show that the echo chambers which recommender systems herd their users into cause epistemic impropriety, no epistemic responsibilities need to be formulated to prevent this effect.

I have three responses. First of all, recommender systems do not promote *one* echo chamber; they promote a variety of such chambers. Every echo chamber has as a central notion that "only those within the chamber are trustworthy". This means that echo chambers promote conflicting beliefs in the domain of trustworthiness judgements. If this is a domain where facts obtain, then this conflict must mean that most – if not all – echo

[186] Jennifer Lackey, "Echo Chambers, Fake News, and Social Epistemology" (Forthcoming), 15–16.

chambers promote false beliefs. Secondly, remember that the one tenet which most social media echo chambers are likely to share, because it helps fulfil the recommender system's goal of attention maximisation, is the idea that "the mainstream media are not to be trusted". But there are, I think, good reasons to doubt this – in fact, I provided such reasons in chapter four. Thus once again, this is a reason to think that social media echo chambers promote false beliefs in the trustworthiness domain – and in doing so, they restrict our access to an epistemically valuable social practice.

Lastly, being part of an echo chamber severely restricts one's access to the knowledge of others in general. This hinders the ability to take advantage of the division of epistemic labour, and therefore to acquire the knowledge one needs to navigate one's life. Baurmann argues similarly, when he says that "(t)he more widespread and the larger the scope of trust networks, the more diverse and detailed the information they aggregate. Particularistic networks which only connect people of a certain category or which are very limited in their scope are constantly in danger of producing misleading, partial and one-sided information."[187] Of course, a restricted trust network of the kind that echo chambers promote can be epistemically beneficial if it is justified: if the majority of people really is not to be trusted, then it is wise to exclude them from one's trust network. But a recommender system does not need an echo chamber to be justified in this way in order to promote it, which means that using a social media service can have considerable negative effects on the reliability of one's trust network.

Together, I think these three points provide sufficient reason to judge social media echo chambers so epistemically undesirable that a responsible epistemic subject ought to do everything in its power to keep itself from being herded into one.

### 3. The epistemic responsibilities of social media services

The argument from the previous section is a *reductio ad absurdum*. It departed from the assumption that we could let users bear all epistemic responsibilities vis-à-vis belief formation on social media services. The conclusions, I submit, are untenable: if the responsibility of users is not to form any beliefs based on what they encounter on social media, then this undercuts the very rationale of these services. Surely – *surely* – news is shared on social media services so that users can form new beliefs about the world; *surely*

[187] Baurmann, "Fundamentalism and Epistemic Authority," 61.

we post eye-witness reports, we participate in online political discussions, we create explanatory videos, so that others can learn from this content. If we want social media services to keep playing the societal roles they currently play, then the propriety of belief-formation through social media services must somehow be salvaged.

The way forward, then, must be to put substantial epistemic responsibilities on the shoulders of social media services themselves. In some cases, the responsibilities can be relatively simple: in order to salvage the knowledge-conduciveness of the news market, it would suffice if social media services only admitted news articles deriving from qualitative sources. Alternatively, social media services could supply reputation information alongside news articles. These suggestions might raise some eyebrows: why should we give social media services the power to influence our trustworthiness judgements? I cannot answer all such criticisms here, but I do wish to point out two things. First, social media services *already* wield substantial power over our cognitive processes: the only differences are that they are currently not open about this, and that they way they wield this power is currently not guided by considerations of epistemic propriety. Second, the responsibility to provide trustworthiness information to users is very similar to the responsibility of traditional newspapers to only employ journalists that fulfil their epistemic responsibilities. Thus although the suggestion might first come across as strange, it is not that far removed from familiar social practices.

Other epistemic responsibilities will be harder to formulate. I am here specifically thinking of the problems caused by recommender systems, which I predominantly discussed in chapters five and six. A first step to improve the epistemic situation must be to provide societal and scientific access to the source codes and datasets of recommender systems. This will allow more in-depth studies of these systems, which will hopefully provide new insights into the epistemic effects they have. This knowledge can then be used to formulate responsibilities in two directions. First, it can be used to inform the digital literacy curriculum: if it is clearer how social media services might be misleading, then it might also become clearer how users can adjust their cognitive processes in order to salvage epistemic propriety. Second, it can be used to construct epistemically responsible recommender systems. The goals in constructing such systems must be twofold: they must aim to restrict to a minimum the ways in which they might affect the reliability of users' cognitive processes; and if they *do* affect this reliability, it must be clear when this happens as well as in which way reliability is affected.

Any suggestion to put substantial epistemic responsibilities on the shoulders of social media services can expect to be met with much resistance. Why should we trust these organisations to fulfil an epistemic role? Is it not better if they remain neutral? Such resistance is intuitive, I think, if we conceive of social media services as epistemically neutral organisations: if they fit in the category of bakeries, swimming pools, and telephone companies, then these services overstep certain boundaries by taking up substantial epistemic responsibilities. But if the arguments in this thesis are correct, then we ought not conceive of social media services as epistemically neutral, but rather as fitting in the category of newspapers, tv-stations, and universities. We ought to conceive of social media services as epistemic organisations, because unlike bakeries, social media services wield distinctly epistemic power. And given that social media services are epistemic organisations, it is a matter of mere epistemic prudence to ensure that they help us meet the standards of epistemic propriety.

# Conclusion

In this thesis, I have investigated the epistemic responsibilities of social media services. My conclusions are not simple: rather than providing a list of responsibilities, I have investigated the nature of epistemic responsibility itself, as well as the epistemic nature of social media services, and have provided some recommendations for directions in which to take the further formulation of the epistemic responsibilities of social media services. My most important conclusions are twofold: social media services are opaque media, and if we want them to keep playing the societal role which they presently play, then distinctly epistemic responsibilities must accrue to these services.

Social media services are opaque, because they affect the reliability of users' cognitive processes without being open about the way reliability is affected. I provided multiple examples to this effect: social media services hinder users' ability to recognise real news; social media services employ intentionally misleading content in order to extend user attention; and the recommender systems which social media services employ might themselves aim to manipulate users' beliefs in order to maximise attention. The opacity is increased because the datasets and source code of recommender systems are kept secret, making it impossible to adequately study which epistemic effects actually obtain.

In order to save users' ability to meet the demands of epistemic propriety, epistemic responsibilities must be distributed over both users and services. If we wish epistemic responsibilities to accrue only to users, then we must swallow a hard pill: lacking epistemic responsibilities on the side of social media services, the responsibility of users will be not to form any beliefs based on what they encounter on social media. This undercuts the role social media services currently play in our societies, making communication by means of social media services impossible.

If we wish to prevent this devastating result, then some epistemic responsibilities must accrue to social media services themselves – or at least, to the people or corporations responsible for the design of these services. The first step in formulating these responsibilities is making public the technical details of the recommender systems that social media services employ, so that we can scientifically study their epistemic effects. Knowing these epistemic effects will allow us to formulate both the responsibilities of users, and the responsibilities of services. In broad lines, the epistemic responsibilities of social media services will be the following: to construct systems that do minimal epistemic harm, and to explicitly communicate the remaining ways in which these systems do such harm.

## References

Ahn, June. “Digital Divides and Social Network Sites: Which Students Participate in Social Media?” *Journal of Educational Computing Research* 45, no. 2 (September 1, 2011): 147–63. https://doi.org/10.2190/EC.45.2.b.

Akester, Sam. “Ludic Loop.” *Sam Akester* (blog), November 11, 2015. https://sammehreviews.wordpress.com/2015/11/11/ludic-loop/.

Alfano, Mark, Amir Ebrahimi Fard, J. Adam Carter, Peter Clutton, and Colin Klein. “Technologically Scaffolded Atypical Cognition: The Case of YouTube’s Recommender System.” *Synthese*, June 9, 2020. https://doi.org/10.1007/s11229-020-02724-x.

AlgoTransparency. “AlgoTransparency Manifesto.” Accessed March 25, 2021. https://algotransparency.org/.

Alter, Adam L., and Daniel M. Oppenheimer. “Uniting the Tribes of Fluency to Form a Metacognitive Nation.” *Personality and Social Psychology Review* 13, no. 3 (August 2009): 219–35. https://doi.org/10.1177/1088868309341564.

Aristotle. *Rhetoric*. South Bend, United States: Infomotions, Inc., 2000. http://ebookcentral.proquest.com/lib/uunl/detail.action?docID=3314386.

Baurmann, Michael. “Fundamentalism and Epistemic Authority.” In *Democracy and Fundamentalism*, edited by A. Aarnio, 45–70. The Tampere Club Series. Tampere: Tampere University Press, 2010.

Bazarova, Natalya N., and Yoon Hyung Choi. “Self-Disclosure in Social Media: Extending the Functional Approach to Disclosure Motivations and Characteristics on Social Network Sites.” *Journal of Communication* 64, no. 4 (August 1, 2014): 635–57. https://doi.org/10.1111/jcom.12106.

Begg, Ian Maynard, Ann Anas, and Suzanne Farinacci. “Dissociation of Processes in Belief: Source Recollection, Statement Familiarity, and the Illusion of Truth.” *Journal of Experimental Psychology: General* 121, no. 4 (1992): 446.

Bergström, Annika, and Maria Jervelycke Belfrage. “News in Social Media: Incidental Consumption and the Role of Opinion Leaders.” *Digital Journalism* 6, no. 5 (May 28, 2018): 583–98. https://doi.org/10.1080/21670811.2018.1423625.

boyd, danah m., and Nicole B. Ellison. “Social Network Sites: Definition, History, and Scholarship.” *Journal of Computer-Mediated Communication* 13, no. 1 (October 2007): 210–30. https://doi.org/10.1111/j.1083-6101.2007.00393.x.

Cain, Caleb. *My Descent into the Alt-Right Pipeline*, 2019. https://www.youtube.com/watch?v=sfLa64_zLrU.

Carroll, Lewis. *Alice's Adventures in Wonderland*. Boston: Lee and Shepard, 1869.

"Center for Humane Technology." Accessed June 16, 2021. https://www.humanetech.com/.

Center for Humane Technology. "How Social Media Hacks Our Brains." Accessed June 12, 2021. https://www.humanetech.com/brain-science.

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. "The Echo Chamber Effect on Social Media." *Proceedings of the National Academy of Sciences* 118, no. 9 (March 2, 2021): e2023301118. https://doi.org/10.1073/pnas.2023301118.

Cohen, Geoffrey L. "Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs." *Journal of Personality and Social Psychology* 85, no. 5 (November 2003): 808–22. https://doi.org/10.1037/0022-3514.85.5.808.

Dallmann, Justin. "When Obstinacy Is a Better (Cognitive) Policy." *Philosophers' Imprint* 17, no. 24 (December 2017): 18.

Davenport, Thomas H., and John C. Beck. *The Attention Economy: Understanding the New Currency of Business*. Boston, Mass: Harvard Business School Press, 2001.

Dechêne, Alice, Christoph Stahl, Jochim Hansen, and Michaela Wänke. "The Truth About the Truth: A Meta-Analytic Review of the Truth Effect." *Personality and Social Psychology Review* 14, no. 2 (May 2010): 238–57. https://doi.org/10.1177/1088868309352251.

Doctorow, Cory. *How to Destroy Surveillance Capitalism*. New York, NY: Stonesong Digital, 2020.

Dormandy, Katherine. "Introduction: An Overview of Trust and Some Key Epistemological Applications." In *Trust in Epistemology*, edited by Katherine Dormandy. New York; London: Routledge, 2020.

———. *Trust in Epistemology*. New York; London: Routledge, 2020.

Engeström, Jyri. "Why Some Social Network Services Work and Others Don't — Or: The Case for Object-Centered Sociality." *Zengestrom* (blog), April 13, 2005. http://www.zengestrom.com/blog/2005/04/why-some-social-network-services-work-and-others-dont-or-the-case-for-object-centered-sociality.html.

Gapper, Paul. “A Satisfying Plot 5: Questions and Puzzles.” *Paulgapper* (blog), June 1, 2014. https://paulgapper.wordpress.com/tag/ludic-loop/.

Gentzkow, Matthew, and Jesse M. Shapiro. “Competition and Truth in the Market for News.” *Journal of Economic Perspectives* 22, no. 2 (2008): 133–54.

Gigerenzer, Gerd, and Wolfgang Gaissmaier. “Heuristic Decision Making.” *Annual Review of Psychology* 62, no. 1 (January 10, 2011): 451–82. https://doi.org/10.1146/annurev-psych-120709-145346.

Goldberg, Sanford. *Assertion: On the Philosophical Significance of Assertoric Speech*. New York, NY: Oxford University Press, 2015.

———. “Epistemic Entitlement and Luck.” *Philosophy and Phenomenological Research* 91, no. 2 (September 2015): 273–302. https://doi.org/10.1111/phpr.12083.

———. “The Division of Epistemic Labor.” *Episteme* 8, no. 1 (February 2011): 112–25. https://doi.org/10.3366/epi.2011.0010.

———. *To the Best of Our Knowledge: Social Expectations and Epistemic Normativity*. First edition. Oxford, United Kingdom: Oxford University Press, 2018.

Goldman, Alvin I. “A Guide to Social Epistemology.” In *Social Epistemology: Essential Readings*, edited by Alvin I. Goldman and Dennis Whitcomb, 11–37. Oxford ; New York: Oxford University Press, 2011.

———. “Experts: Which Ones Should You Trust?” *Philosophy and Phenomenological Research* 63, no. 1 (2001): 85–110.

———. *Knowledge in a Social World*. Reprint. Oxford: Clarendon Press, 2003.

Goldman, Alvin I., and James C. Cox. “Speech, Truth, and the Free Market for Ideas.” *Legal Theory* 2, no. 1 (1996): 1–32.

Harris, Tristan, and Aza Raskin. “Your Undivided Attention,” n.d. https://www.humanetech.com/podcast.

Heaven, Douglas. “Engineered Compulsion: Why Candy Crush Is the Future of More than Games.” *New Scientist* 222, no. 2971 (May 2014): 38–41. https://doi.org/10.1016/S0262-4079(14)61069-1.

Herrman, John. “How TikTok Is Rewriting the World.” *The New York Times*, March 10, 2019. https://www.nytimes.com/2019/03/10/style/what-is-tik-tok.html.

Hill, Robin K. “What an Algorithm Is.” *Philosophy & Technology* 29, no. 1 (March 2016): 35–59. https://doi.org/10.1007/s13347-014-0184-5.

"How to Earn Money on YouTube - YouTube Help." Accessed June 12, 2021. https://support.google.com/youtube/answer/72857?hl=en.

Facebook for Business. "How to Make Money from Your Content on Facebook." Accessed June 12, 2021. https://en-gb.facebook.com/business/learn/lessons/how-make-money-facebook.

James, Richard J. E., Claire O'Malley, and Richard J. Tunney. "Understanding the Psychology of Mobile Gambling: A Behavioural Synthesis." *British Journal of Psychology* 108, no. 3 (2017): 608–25. https://doi.org/10.1111/bjop.12226.

Jamieson, Kathleen Hall, and Joseph N Cappella. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford; New York: Oxford University Press, 2010.

Jannach, Dietmar, and Gediminas Adomavicius. "Recommendations with a Purpose." In *Proceedings of the 10th ACM Conference on Recommender Systems*, 7–10. Boston Massachusetts USA: ACM, 2016. https://doi.org/10.1145/2959100.2959186.

Jaster, Romy, and David Lanius. "Schlechte Nachrichten: >>Fake News<< in Politik Und Öffentlichkeit." In *Fake News Und Desinformation: Herausforderungen Für Die Vernetzte Gesellschaft Und Die Empirische Forschung*, edited by Ralf Hohlfeld, Michael Harnischmacher, Elfi Heinke, Lea Lehner, and Michael Sengl. Nomos Verlagsgesellschaft mbH & Co. KG, 2020. https://doi.org/10.5771/9783748901334.

Kahan, Dan M. "On the Sources of Ordinary Science Knowledge and Extraordinary Science Ignorance." In *Oxford Handbook on the Science of Science Communication*, edited by Kathleen Hall Jamieson, Dan M. Kahan, and Dietram A. Scheufele, Vol. 1. Oxford University Press, 2017. https://doi.org/10.1093/oxfordhb/9780190497620.013.4.

Kietzmann, Jan H., Kristopher Hermkens, Ian P. McCarthy, and Bruno S. Silvestre. "Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media." *Business Horizons* 54, no. 3 (May 2011): 241–51. https://doi.org/10.1016/j.bushor.2011.01.005.

Kräenbring, Jona, Tika Monzon Penza, Joanna Gutmann, Susanne Muehlich, Oliver Zolk, Leszek Wojnowski, Renke Maas, Stefan Engelhardt, and Antonio Sarikas. "Accuracy and Completeness of Drug Information in Wikipedia: A Comparison with Standard Textbooks of Pharmacology." *PloS One* 9, no. 9 (2014): e106930.

Lackey, Jennifer. “Echo Chambers, Fake News, and Social Epistemology,” Forthcoming.

Levy, Neil. “The Bad News about Fake News.” *Social Epistemology Review and Reply Collective* 6, no. 8 (2017): 20–36.

Levy, Ro’ee. “Social Media, News Consumption, and Polarization: Evidence from a Field Experiment.” *American Economic Review* 111, no. 3 (March 1, 2021): 831–70. https://doi.org/10.1257/aer.20191777.

Lu, Xing, Zhicong Lu, and Changqing Liu. “Exploring TikTok Use and Non-Use Practices and Experiences in China.” In *Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing*, edited by Gabriele Meiselwitz, 57–70. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-49576-3_5.

Luca, Michael. “Social Media Bans Are Really, Actually, Shockingly Common.” WIRED, January 20, 2021. https://www.wired.com/story/opinion-social-media-bans-are-really-actually-shockingly-common/.

Mathiesen, Kay. “Fake News and the Limits of Freedom of Speech.” In *Media Ethics, Free Speech, and the Requirements of Democracy*, edited by Carl Fox and Joe Saunders, 161–79. Routledge Research in Applied Ethics 13. New York: Routledge, 2019.

McGlone, Matthew S., and Jessica Tofighbakhsh. “Birds of a Feather Flock Conjointly (?): Rhyme as Reason in Aphorisms.” *Psychological Science* 11, no. 5 (September 1, 2000): 424–28. https://doi.org/10.1111/1467-9280.00282.

Meyerson, Eric. “YouTube Now: Why We Focus on Watch Time.” *Blog.Youtube* (blog). Accessed March 3, 2021. https://blog.youtube/news-and-events/youtube-now-why-we-focus-on-watch-time/.

Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi. “Recommender Systems and Their Ethical Challenges.” *AI & SOCIETY* 35, no. 4 (December 2020): 957–67. https://doi.org/10.1007/s00146-020-00950-y.

Moran, Richard. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton, New Jersey: Princeton University Press, 2001.

Morton, Adam. “Human Bounds: Rationality for Our Species.” *Synthese* 176, no. 1 (September 2010): 5–21. https://doi.org/10.1007/s11229-009-9481-4.

Mukerji, Nikil. “What Is Fake News?” *Ergo, an Open Access Journal of Philosophy* 5, no. 20201214 (December 11, 2018): 923–46. https://doi.org/10.3998/ergo.12405314.0005.035.

Nakamura, Jeanne, and Mihaly Csikszentmihalyi. “The Concept of Flow.” In *Handbook of Positive Psychology*, edited by C.R. Snyder and S.J. Lopez, 89–105. Oxford: Oxford University Press, 2005. https://doi.org/10.1007/978-94-017-9088-8_16.

Newman, Eryn. J., Madeline C. Jalbert, Norbert Schwarz, and Deva P. Ly. “Truthiness, the Illusory Truth Effect, and the Role of Need for Cognition.” *Consciousness and Cognition* 78 (February 2020): 102866. https://doi.org/10.1016/j.concog.2019.102866.

Nguyen, C. Thi. “Echo Chambers and Epistemic Bubbles.” *Episteme* 17, no. 2 (2020): 141–61. https://doi.org/10.1017/epi.2018.32.

———. “The Seductions of Clarity.” *Royal Institute of Philosophy Supplement* 89 (May 2021): 227–55. https://doi.org/10.1017/S1358246121000035.

Obar, Jonathan A., and Steve Wildman. “Social Media Definition and the Governance Challenge: An Introduction to the Special Issue.” *Telecommunications Policy* 39, no. 9 (October 2015): 745–50. https://doi.org/10.1016/j.telpol.2015.07.014.

O’Callaghan, Derek, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. “Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems.” *Social Science Computer Review* 33, no. 4 (August 2015): 459–78. https://doi.org/10.1177/0894439314555329.

O’Connor, Cailin, and James Owen Weatherall. “Modeling How False Beliefs Spread.” In *The Routledge Handbook of Political Epistemology*, edited by Michael Hannon and Jeroen de Ridder, 203–13. Routledge Handbooks in Philosophy. Abingdon, Oxon ; New York, NY: Routledge, 2021.

Orlowski, Jeff. *The Social Dilemma*. Docudrama. Netflix, 2020.

Peirce, Charles Sanders. “The Fixation of Belief.” In *Classics of Western Philosophy*, edited by Steven M. Cahn, 8th ed., 1246–54. Indianapolis/Cambridge: Hackett Publishing Company, Inc., 2012.

Pettit, Philip. “The Conversable, Responsible Corporation.” In *The Moral Responsibility of Firms*, edited by Eric W. Orts and N. Craig Smith, 15–35. Oxford University Press, 2017. https://doi.org/10.1093/oso/9780198738534.003.0002.

Reber, Rolf, and Christian Unkelbach. “The Epistemic Status of Processing Fluency as Source for Judgments of Truth.” *Review of Philosophy and Psychology* 1, no. 4 (December 2010): 563–81. https://doi.org/10.1007/s13164-010-0039-7.

Relman, Eliza. “Right-Wing Media Has Pushed 3 Completely False Narratives in Less than a Week.” Business Insider, April 27, 2021. https://www.businessinsider.com/right-wing-media-fox-news-3-debunked-stories-in-week-2021-4.

Ricci, Francesco, Lior Rokach, and Bracha Shapira. “Recommender Systems: Introduction and Challenges.” In *Recommender Systems Handbook*, edited by Francesco Ricci, Lior Rokach, and Bracha Shapira, 1–34. Boston, MA: Springer US, 2015. https://doi.org/10.1007/978-1-4899-7637-6_1.

Rini, Regina. “Fake News and Partisan Epistemology.” *Kennedy Institute of Ethics Journal* 27, no. 2S (2017): E-43-E-64. https://doi.org/10.1353/ken.2017.0025.

Robeyns, Ingrid. “Ideal Theory in Theory and Practice:” *Social Theory and Practice* 34, no. 3 (2008): 341–62. https://doi.org/10.5840/soctheorpract200834321.

Roose, Kevin. “Rabbit Hole.” Accessed June 1, 2021. https://www.nytimes.com/column/rabbit-hole.

Ryan, Shane. “Epistemic Environmentalism:” *Journal of Philosophical Research* 43 (2018): 97–112. https://doi.org/10.5840/jpr201872121.

Schmitt, Frederick F., and Reza Lahroodi. “The Epistemic Value of Curiosity.” *Educational Theory* 58, no. 2 (May 2008): 125–48. https://doi.org/10.1111/j.1741-5446.2008.00281.x.

Schüll, Natasha Dow. *Addiction by Design: Machine Gambling in Las Vegas*. Princeton, NJ: Princeton University Press, 2012.

———. NPR All Things Considered. Interview by Arun Rath. Radio, June 7, 2014.

Silverman, Craig. “This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook.” BuzzFeed News, November 16, 2016. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

Simon, Herbert A. “Designing Organizations for an Information-Rich World.” In *Computers, Communications, and the Public Interest*, 38–52. Baltimore, MD: The John Hopkins Press, 1971.

Sunstein, Cass R. *#Republic: Divided Democracy in the Age of Social Media*. Princeton ; Oxford: Princeton University Press, 2017.

Tandoc, Edson C., Zheng Wei Lim, and Richard Ling. “Defining ‘Fake News’: A Typology of Scholarly Definitions.” *Digital Journalism* 6, no. 2 (February 7, 2018): 137–53. https://doi.org/10.1080/21670811.2017.1360143.

American Press Institute. “The Elements of Journalism.” Accessed April 1, 2021. https://www.americanpressinstitute.org/journalism-essentials/what-is-journalism/elements-journalism/.

Tsamados, Andreas, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. “The Ethics of Algorithms: Key Problems and Solutions.” *AI & SOCIETY*, February 20, 2021. https://doi.org/10.1007/s00146-021-01154-8.

Tufekci, Zeynep. “YouTube, the Great Radicalizer.” *The New York Times* 10 (March 10, 2018).

United States Government. “Transcript of Mark Zuckerberg’s Senate Hearing.” *Washington Post*, April 11, 2018. https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/.

Urban Dictionary. “Urban Dictionary: Rabbit Hole,” August 19, 2013. https://www.urbandictionary.com/define.php?term=Rabbit%20Hole.

Urban Dictionary. “Urban Dictionary: Rabbit Hole,” June 17, 2015. https://www.urbandictionary.com/define.php?term=Rabbit%20Hole.

Van den Bulck, H, and A Hyzen. “Of Lizards and Ideological Entrepreneurs: Alex Jones and Infowars in the Relationship between Populist Nationalism and the Post-Global Media Ecology.” *International Communication Gazette* 82, no. 1 (February 2020): 42–59. https://doi.org/10.1177/1748048519880726.

Weatherall, James Owen, Cailin O’Connor, and Justin P. Bruner. “How to Beat Science and Influence People: Policymakers and Propaganda in Epistemic Networks.” *The British Journal for the Philosophy of Science* 71, no. 4 (December 1, 2020): 1157–86. https://doi.org/10.1093/bjps/axy062.

Weaver, Kimberlee, Stephen M. Garcia, Norbert Schwarz, and Dale T. Miller. “Inferring the Popularity of an Opinion from Its Familiarity: A Repetitive Voice Can Sound like a Chorus.” *Journal of Personality and Social Psychology* 92, no. 5 (May 2007): 821–33. https://doi.org/10.1037/0022-3514.92.5.821.

"Wiki Rabbit Hole." In *Wikipedia*, April 2, 2021. https://en.wikipedia.org/w/index.php?title=Wiki_rabbit_hole&oldid=1015623747.

Williams, Bernard. *Truth & Truthfulness: An Essay in Genealogy*. Princeton, N.J: Princeton University Press, 2002.

Williamson, Elizabeth, and Emily Steel. "Conspiracy Theories Made Alex Jones Very Rich. They May Bring Him Down." *The New York Times*, September 7, 2018, sec. U.S. https://www.nytimes.com/2018/09/07/us/politics/alex-jones-business-infowars-conspiracy.html.

Zarley, B David. "Inside the Fight to Make Tech More Humane." A Beautiful Perspective, March 2, 2018. https://abeautifulperspective.com/2018/03/inside-the-fight-to-make-tech-more-humane/.

Zollman, Kevin J. S. "The Communication Structure of Epistemic Communities." *Philosophy of Science* 74, no. 5 (December 2007): 574–87. https://doi.org/10.1086/525605.

Zuckerman, Ethan, and Chand Rajendra-Nicolucci. "Beyond Facebook Logic: Help Us Map Alternative Social Media!," October 8, 2020. https://knightcolumbia.org./content/beyond-facebook-logic-help-us-map-alternative-social-media.