Mapping Value(s) in AI: Methodological Directions for Examining Normativity in Complex Technical Systems

Rieder, B.; Gordon, G.; Sileno, G.

Mapping Value(s) in AI: Methodological Directions for Examining Normativity in Complex Technical Systems

Bernhard Rieder[*] [a]   Geoff Gordon [b]   Giovanni Sileno [c]

[a]   Department of Media Studies, University of Amsterdam (The Netherlands)
[b]   Asser Institute (The Netherlands)
[c]   Informatics Institute, University of Amsterdam (The Netherlands)

Abstract

This paper seeks to develop a multidisciplinary methodological framework and research agenda for studying the broad array of 'ideas', 'norms', or 'values' incorporated and mobilized in systems relying on AI components. We focus on recommender systems as a broader field of technical practice and take YouTube as an example of a concrete artifact that raises many social concerns. To situate the conceptual perspective and rationale informing our approach, we briefly discuss investigations into normativity in technology more broadly and refer to 'descriptive ethics' and 'ethigraphy' as two approaches concerned with the empirical study of values and norms. Drawing on science and technology studies, we argue that normativity cannot be reduced to ethics, but requires paying attention to a wider range of elements, including the performativity of material objects themselves. The method of 'encircling' is presented as a way to deal with both the secrecy surrounding many commercial systems and the socio-technical and distributed character of normativity more broadly. The resulting investigation aims to draw from a series of approaches and methods to construct a much wider picture than what could result from one discipline only. The paper is then dedicated to developing this methodological framework organized into three layers that demarcate specific avenues for conceptual reflection and empirical research, moving from the more general to the more concrete: ambient technical knowledge, local design conditions, and materialized values. We conclude by arguing that deontological approaches to normativity in AI need to take into account the many different ways norms and values are embedded in technical systems.

Keywords: Normativity; machine learning; empirical ethics; YouTube; methodology.

*   ✉ b.rieder@uva.nl

1   Introduction

Over the past decade, complex algorithmic systems have proliferated while also coming increasingly under critical scrutiny. Terms like artificial intelligence, big data, or machine learning designate vectors of technological change that are often seen as transformative but at the same time as opaque and elusive, in part due to their technical complexities (Burrell, 2016; Pasquale, 2015). Particularly in the context of large online platforms, there is a stark contrast between anxieties about potentially far-reaching effects of ranking, recommendation, classification, personalized targeting, and other forms of "artificial cognition" on the one side and the considerable impediments to getting a firm understanding of the mechanisms behind these functions on the other. The fields of Explainable Artificial Intelligence and, more specifically, Explainable Machine Learning are dedicated to answering the question "how do we understand the decisions suggested by these systems in order that we can trust them?" (Belle & Papantonis, 2020, p. 1), but the methods put forward generally need to be implemented as part of the system, requiring the collaboration and good will of its creators. Given the culture of secrecy surrounding corporate technology projects, implementing such methods and making their results available for public scrutiny will prove difficult in most cases. Furthermore, approaches focusing on the decision models at the core of machine learning systems often rely on definitions of "explanation" and "understanding" primarily addressing "interpretability" concerns that leave out much of what surrounds these models, including the question of why there needs to be an AI system at all (Powles & Nissenbaum, 2018).

Another way to interpret and assess the behavior of a technical system is to inquire more broadly into its *design*, that is, into how and why it was made, with the awareness that the process of its making will have involved choices, presuppositions, and normative commitments that influence how it will perform. While there are long-standing research traditions sensitive to "values in design", the question of how norms and values are embedded in complex computing systems operating within specific social contexts remains challenging. This is in part due to the fraught problem of choosing between competing or mutually exclusive norms and values[1], but also because many design decisions are seemingly innocuous, and their ramifications are not obvious on first analysis. The task of most AI systems, however, is to *differentiate* — and even "intelligent" differentiations more often than not designate winners and losers, for example when some person or piece of content is given preference over another.

Taking a recommender system operating in a large and tight-lipped online platform as its main example, this paper seeks to develop a multidisciplinary methodological toolbox and framework anchored in the humanities and social sciences, but also takes cues from law and computer science, to study the broad array of "ideas", "norms", or "values" such a system may incorporate and mobilize. But what do these terms mean in the context of technical artifacts? Are they simple transpositions of non-technical constructs into technical form? Instead of limiting ourselves to a narrow understanding of normativity, we situate algorithmic systems within complex networks of "value(s)" that shape the production of technical objects and, consequently, the objects themselves, for example, translations from economic models, social theories, legal requirements, ethical principles, technical knowledge, experiential evaluations, or other constructs used to define and justify design goals and decisions. We conflate value and values, here, to signal their common association with desirability and to acknowledge the entanglement between economic matters and other normative concerns.

---

1.   This problem manifests, for example, around the mathematical incommensurability of different value definitions (e.g. Kleinberg et al., 2016).

Technical systems are designed artifacts: they incorporate commitments to a set of prescriptions when they are built, and they enact and reproduce these commitments when they are operating. But how can we account for the different strands of normativity that coagulate around AI-based systems more broadly and around specific systems in particular? Normativity, on the broadest level, refers to the phenomenon that individuals or groups designate some things as desirable and others undesirable. We use the term, here, to put the spotlight in the first instance onto the various streams of *reasoning* that inform how a technical system *ought* to look and perform. How does a recommender system come to define what constitutes a "good" recommendation? But given the specificities of contemporary techniques like machine learning and the continuous and distributed nature of agency within large online platforms, the term "reasoning" is also limiting, since it does not capture non-intended and emerging properties or behaviors. When we ask why a system behaves like it does, we can try to understand its designers' beliefs and intentions or analyze the incentives and legal pressures they operate under, but we also have to recognize that part of its normative thrust — the way it intervenes and makes judgements in concrete situations — emerges from the dynamic coupling with its environment (Rahwan et al., 2019). Values, as we understand them, are thus not merely idealized presuppositions that inform a technical system from the outside, imprinting motives and ethics onto an otherwise neutral object, but something that is itself produced in and through the wider ecologies that system is embedded in.[2] There is a need for a broader articulation of normativity that connects values to performativity, keyed to the actual "work" done by computational artifacts.

Rather than asking which values AI *should* satisfy or discussing ethical challenges (Milano, 2019), this paper tackles the problem of how to study different forms of normativity as they run through engineering traditions, concrete application contexts, and actual technical artifacts. To have tangible reference points, we focus on recommender systems as a broader field of technical practice and take YouTube as an example of a concrete artifact that resists "explainability" from the outside, yet clearly raises a variety of social concerns, including radicalization and polarization (e.g. Yesilada & Lewandowsky, 2022), misinformation (e.g. Li et al., 2020), and cultural gatekeeping (e.g. Bonini & Gandini, 2019). The goal is not to provide a full-fledged analysis of the various instances of normativity that play out in these contexts, but to tie different methodological directions to a joint progression over different layers. As Bucher (2018) argues, "[w]hile we cannot ask the algorithm in the same way we may ask humans about their beliefs and values, we may indeed attempt to find other ways of making it 'speak'" (p. 60). What follows is thus an attempt to assemble a discipline-spanning research program for making complex algorithmic systems operating in corporate settings "speak" in a number of different ways, paying particular attention to background conditions, design contexts, and materiality. In contrast to the model-focused methods we find in Explainable Machine Learning, we situate these elements within and in relation to an emerging "platform society" (van Dijck et al., 2018), where algorithmic mechanisms are tied up with commodification and business models. In line with other critiques of the "black box" metaphor (e.g. Christin, 2020; Straube, 2019), we believe that even if we were to get access to source code or trained models, there would be much left to explain, in particular when it comes to understanding the reasons behind design decisions, evaluation procedures, or emergent properties. Our distinctive contribution is thus to assemble and connect approaches that focus specifically on the relationship between such normative constructs and concrete instances of technology. Too often confined to their specific fields, we consider that these approaches are complementary and that bringing them together

---

2. We develop this argument in more detail in Gordon et al. (2022).

allows us to paint a more holistic picture of complex technologies operating in contexts that resist inquiry in a variety of ways.

The paper is structured as follows. Section 2 further elaborates on the rationale behind our approach and proposed methodology, inspired by the technique of "encircling" developed in security studies by de Goede et al. (2019) and recalibrated here for the study of complex socio-technical systems. Over sections 3, 4, and 5, we simultaneously sketch out individual methodological components and "apply" them through a number of "analytical stubs": we discuss recommender systems and YouTube's specific case through three analytical layers that "circle" progressively closer to the technical object itself, addressing, in turn, ambient technical knowledge, local design conditions, and materialized values.

2   Rationale

Scholars have long argued that agency (Latour, 2005) and cognition (Hutchins, 1995) are distributed over human and non-human components, opening the door for the conceptual and empirical inclusion of "things" as carriers or agents of normativity. Some have asked even more explicitly whether "artifacts have politics" (Winner, 1980). In the context of computer systems, the field known as "values in design" (Nissenbaum, 1998; Knobel & Bowker, 2011) stands out. Already since the early 1990s, it combines an interest in normative concerns such as bias, autonomy, or privacy with more proactive attempts to integrate the awareness and sensitivity for these concerns within design processes and methodologies themselves. Some of this work has found expression in critiques of web search engines, which raise questions that are in many ways similar to those surrounding more modern-sounding terms like artificial intelligence, machine learning, or algorithmic decision-making. Search engines parse very large pools of data with the help of algorithms in ways that "dictate systematic prominence for some sites, dictating systematic invisibility for others" (Introna & Nissenbaum, 2000, p. 171) and their "politics" constitute a normative assertion, albeit expressed as mechanism rather than principle.

Since then, rampant computerization, datafication, and the emergence of techniques like deep learning have widened application spaces to almost any kind of *ordering* (ranking, prioritizing, deciding, etc.) online. This includes content recommendation and advertisement as well as socially sensitive domains such as criminal justice, hiring, or access to credit. Work coming from the humanities and social sciences has been particularly concerned with (potential) effects on individuals, specific groups, or society at large (e.g. O'Neil, 2016; Eubanks, 2018) and often identifies the inherent opacity of machine learning techniques (e.g. Burrell, 2016) as a major roadblock to understanding and critique.

The already mentioned efforts in computer science to create "explainable" algorithms that are capable of providing human-understandable reasoning for specific decisions have been met by attempts to promote algorithmic accountability from the "outside" through auditing, reverse-engineering, and other techniques (e.g. Diakopoulos, 2015; Sandvig et al., 2014). These approaches localize the normative thrust of a larger system within contained technical procedures and inquire into forms of bias or epistemological character, sometimes in great detail. But when it comes to defining the broader normative horizon such systems are embedded in, we often find little more than referrals to broad narratives, such as presumed beliefs in technical objectivity or the imperatives of "surveillance capitalism" (Zuboff, 2019).

At the same time, we are witnessing a frantic proliferation of deontological principles targeting technical objects and engineering practices. Professional societies like the ACM (Association for Computing Machinery, 2018) and IEEE (Institute of Electrical and Electronics

Engineers, n.a.) have published "codes of conduct" and similar guidelines for decades, but we are now increasingly seeing proposals covering the various fields engaging with complex algorithms more specifically. The ACM's (2017) *Statement on Algorithmic Transparency and Accountability* or the EU's *Ethics Guidelines for Trustworthy AI* (European Commission, 2019) are two examples from the 84 ethics documents surveyed in Jobin et al. (2019). While these reflections and guidelines are clearly important, they cannot answer the more empirical question of knowing how and which values *are* being embedded and activated in concrete settings and systems. And this includes new value constructions enabled by complex technical systems, not solely the pre-existing values incorporated into them.

This task falls to fields like descriptive ethics, which tackles "the challenge of providing rich and accurate pictures of the moral conditions, values, virtues, and norms, under which people live and have lived" (Hämäläinen, 2016, p. 1). A full account of the many elements of normativity involved in the creation, maintenance, and further development of complex technical objects is of course an impossibility and any attempt at creating a "rich and accurate picture" is necessarily partial. On the broadest level, technical creation is embedded in what Gert (2004) calls "common morality", an almost universal baseline that includes avoidance of five key elements: death, pain, disability, loss of freedom, and loss of pleasure. But below these generalities, we encounter an intricate landscape of values and norms, unfolding in segmented and stratified societies marked by conflict and competing interests.

Similar to the goals of descriptive ethics, we seek to submit our case to something akin to what Lynch (2001) calls an "ethigraphy", an "empirical ethics" that assesses how moral agents act and reason in circumscribed situations. In his work on search engine optimization practices, Ziewitz (2019) expands on Lynch and defines "ethics as a practical accomplishment" (p. 714) rather than a disembodied set of rules. The contents and processes behind this "ethical work" (p. 713) is then made the subject of empirical study. But unlike Seaver (2017), who was able to gain access to a smaller recommender system company for his own ethnographic work, systems like YouTube's are hardly open to this type of investigation. In fact, the reticence of large online platforms to communicate around the inner workings of their algorithmic systems has been a common complaint (e.g. van Dijck et al., 2018; Rieder & Hofmann, 2020), to the point where regulations such as the EU's Digital Services Act (European Commission, 2020) are trying to establish "transparency obligations" with regard to ranking and recommendation.

We thus take inspiration from the technique of "encircling", recently developed by de Goede and colleagues (2019) in the context of security studies, where secrecy and obfuscation are a constant reality. While there are fundamental differences between recommender systems operating in online platforms on the one hand, and, on the other hand, things like military operations, public and private security institutions and practices, international relations, arms trading, secret prisons, criminal organizations, and other subjects security researchers commonly deal with, all of these things share a setting where "public disclosure is the exception and secrecy the norm" (Walters, 2014, as cited in de Goede et al., 2019, p. 5). Since various kinds of barriers may thwart access to information in such settings, de Goede et al. (2019) advocate a spirit of improvisation and *bricolage*:

> Encircling entails a lateral, multipronged, creative, iterative approaching of secret sites, confidential materials and classified practices. It is less focused on uncovering the kernel of the secret, than it is on analysing the mundane lifeworlds of security practices and practitioners that are powerfully structured through codes and rites of secrecy. (p. 14)

Similar to critiques of algorithmic accountability as "opening the black box", the goal is not merely to establish some kind of factual truth, but to make secrecy itself productive to the analysis: "mapping secrecies and sensitivities in the field can itself be revealing; navigating obfuscation is co-productive of research design and data" (de Goede et al., p. 7). For our project, this does not mean giving up on the idea that something can be said about the workings of a recommender system without access to insider information. Rather, it means that its creators' choice to communicate about certain aspects and not others is itself a normative act that can be seen both as a finding and as a performative gesture that affects how users, researchers, regulators, and other actors position themselves with regard to the system.

The parallel with security practices also suggests that researchers must draw on a wide array of methods, materials, informants, and opportunities that may vary substantially from case to case. This is not only a concession to secrecy and lack of access, but also linked to a broader understanding of the distributed nature of complex socio-technical settings. Concrete systems are indeed best described as complex *assemblages* that combine many different technical components with various kinds of social embeddings (Ananny & Crawford, 2016). We use the term "technical system" rather than "algorithm" to highlight that normativity and normative concerns are not limited to what we call "AI components" but distributed over larger chains or networks. Data collection and preparation processes[3], for example, are highly important for how a system behaves and interface arrangements mediate how end-users experience outputs. But more "organizational" processes are equally crucial, for example when preferential constructs for a machine learning module (e.g. optimization targets) are decided upon and associated control activities (e.g. performance evaluation procedures) are put into place. Ultimately, technical systems and their social contexts form a continuum rather than two separate categories, and accounting for the behavior of a concrete technical system will have to draw on a variety of explanations.

When thinking in terms of encircling, the goal is therefore not to "unmask" a hidden agenda but rather to reconstruct and assess how different nodal points in the overall assemblage are connected with normative effect whenever such a system is designed, implemented, and developed further over time. This includes explicit normative decisions, implicit beliefs, and "practical" values, for example those running through specific engineering traditions, as well as adaptations that emerge as the system is used, which may be unintended but nonetheless affect people's behavior and opportunities. Values and the work they do in concrete settings are enacted in practices and reified in complex artifacts. This also means that focusing on the "ethics" of practitioners in the narrow sense is not enough. From an actor-network perspective (Latour, 2005) in particular, value construction can be understood as co-substantial with *agency* and therefore as part of almost anything. The question, then, is where to begin and where to end. Encircling proposes to approach this question inductively and iteratively, which means that central sites or aspects can be identified and "encircled" progressively as our understanding of a particular case grows. While envisaging norms and values as distributed over infinitesimal instances of agency necessarily destabilizes both our conceptual and methodological grasp, encircling becomes a way of "re-assembling" (Latour, 2005) our terrain into a more coherent narrative, even if the object remains indeterminate in fundamental ways. Encircling ultimately allows us to acknowledge and maintain this indeterminacy, relatively reducing it by putting a boundary around it but without claiming to define the exact center.

---

3. For instance, many of the techniques proposed to implement "fair AI" consist in adequately selecting or adjusting the training data set (Friedler et al., 2019).

In our case, we set the boundaries of investigation in the first instance by taking a specific technical field — recommender systems — as our main focus. While recommender systems are not fundamentally different from other information ordering devices in technical terms, they present a use case that does not follow the canonical query-result format. Relying most often on indirect user input (i.e. logged behavior), they encapsulate ideas about what constitutes a "good" recommendation that are less amenable to user intervention and therefore particularly interesting for a perspective concerned with normative analysis. Our goal, here, is to better understand how "quality" or "success" are being defined within the field, how practical, operational, epistemological, and normative concerns intertwine. And we are notably interested in the broad range of considerations that come into play, not merely in explicitly "ethical" criteria like bias, which is the most common way of discussing the normative thrust of AI (Pasquinelli, 2019). In the second instance, we aim to bridge the gap between the more general category of recommender systems and concrete implementations that have "real-world" implications. Taking YouTube as our main example indeed allows us to connect with pressing contemporary concerns — and with a quickly growing field of research that we draw on throughout the paper. As already mentioned, how the platform recommends videos to users has come under scrutiny with regard to radicalization and misinformation (e.g. Yesilada & Lewandowsky, 2022; Li et al., 2020), but there have also been critical interrogations concerning the pressures it puts on creators (e.g. Bishop, 2019; Kumar, 2019) and broader takes on cultural gatekeeping (e.g. Bonini & Gandini, 2019). Since YouTube is relatively accommodating when it comes to data collection via scraping, there have been a number of empirical studies targeting the behavior of the recommendation system directly (e.g. Airoldi et al., 2016; Ribeiro et al., 2019; Alfano et al., 2020 Matamoros-Fernández et al., 2021), which we will come back to in section five. The considerable attention AI-fueled video recommendations have already received is not only justified by the overall size and global reach of the platform, but also by YouTube Chief Product Officer Neal Mohan's revelation that over 70% of video watchtime is driven by these systems (Solsman, 2018).

2.1 Methodology overview

In line with the spirit of improvisation and iteration recommended by de Goede et al. (2019), we consider the eclectic mix of approaches and methods that follows not as a fixed or complete recipe for the investigation of how normativity informs complex technical objects. The different ingredients have been encountered and chosen *inductively*, from extensive literature review and our own situated experiences with studying AI systems from our disciplinary backgrounds in media studies, law, and computer science. As Straube argues in his attempt to adapt social scientific research to the study of digital devices such as algorithms, "the most helpful methodological toolkit depends heavily on the precise context of the research and the affordances presenting themselves in the field" (2019, p. 188). In our specific case, and for corporate owned- and operated AI systems in general, the toolkit is heavily predicated towards methods that do not require access to actors working at the companies in question. While we do discuss ethnographic approaches in chapter four, this is not an attempt to adapt traditions like multi-sited ethnography (Marcus, 1995), as Bosma (2019) does in her work on digital security technologies, which still rely mainly on interviewing and observing the (human) actors assembling around these systems. Instead, we take inspiration from projects like Eriksson et al.'s (2019) in-depth study of Spotify, which faced considerable resistance from the company, to think about how to complement "front-end inquiries (such as interviewing) with experi-

mental back-end studies of digital media infrastructure, metadata generation, and aggregation practices" (p. 2). Their opportunity-driven approach, which embraced unpredictability and combined a variety of methodological "interventions", including setting up a record label and intercepting network traffic, is indeed a good example for what we have in mind.

Our proposal is thus both more limited and more expansive than "traditional" science and technology studies. More limited, since we are specifically interested in articulations of normativity that emerge around and within technical systems, which help us explain and account for the behavior of these systems. This comes at the price of losing other dimensions of analysis, but allows us to connect more directly to the normative debates that surround AI, not least questions of regulation. The goal, eventually, is to *critique* as well as analyze. We also leave to the side "effects", such as social impacts, focusing on what goes *into* the machine, again to understand how and why it does what it does. More expansive, because we are specifically interested in the knowledge traditions and broader modes of reasoning that inform the work on recommender systems in general, across sites and cases, before concrete systems come into play. And, on the other end, we consider concrete systems themselves to be central "informants" about the normative commitments they incorporate, foregrounding methods that make their behavior amenable to critical examination.

While the specific methods mix has to be calibrated to the case at hand, the last two points suggest a more general setup, organized around three layers, that has broader applicability. Informing both conceptual reflection and empirical research, we move from the more general to the more concrete and distinguish ambient technical knowledge, local design conditions, and materialized values.

Ambient technical knowledge addresses the "archives" and traditions technologists draw on when designing a system; the field of recommender systems has a history and a "substance" that includes technical knowledge, best practices, evaluative procedures, and readily available code. As Agre (1997b) argues, computer science overall and specific subfields in particular come with their own "cultural forms", that is, with "linguistic forms, habits of thought, established techniques, ritualized work practices, ways of framing questions and answers, genre conventions, and so forth" (p. 150), which technical practitioners are not necessarily fully aware of. The engineers and scientists working on YouTube's recommender systems enter the company not as blank slates, but are already immersed in cultures of knowledge and practice. In addition to providing the first normative cues, reconstructing these cultural forms has two other objectives: on the one hand, it helps researchers from non-technical disciplines acquire some level of technical knowledge, combating what Van Veeren (2018) calls "invisibility as inexpertise"; on the other hand, it provides elements and dimensions to look for when analyzing what comes and, crucially, what comes *not* into play in concrete production settings. *Local design conditions* then stand for the particular circumstances and considerations that influence how a concrete system is designed. While access to companies' internal reasoning is often not an option, there are other materials, publications, and informants to draw on to further understand the choices and normative commitments made in the construction of a technical object. Analyzing the economic and legal context can inform on incentives and other conditioning forces that weigh on these choices. Attending to *materialized values*, finally, responds directly to the call "to complement ethnographic approaches with novel methods to observe technology" (de Goede et al., 2019, p. 15). It addresses technical systems themselves as normative agents, in the sense that they act and provide opportunities for action in ways that are directed and prescriptive. Contemporary systems' internals may be opaque, but they are "observable" (Rieder & Hofmann, 2020) to a certain degree, affording the opportunity to investigate their normative thrust by

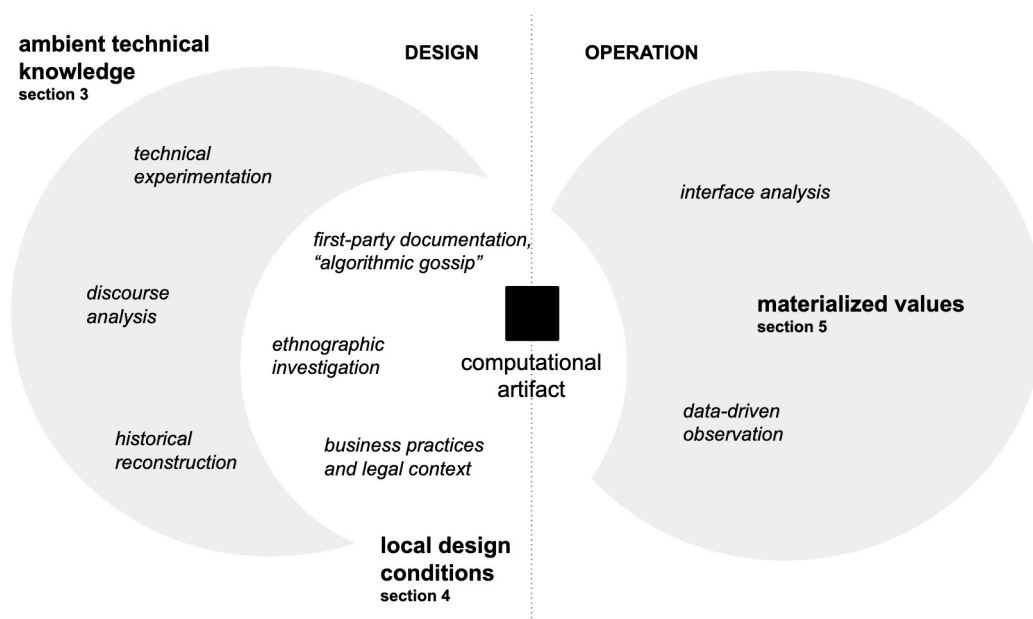analyzing their interfaces and technical operations.

Figure 1: Overview of the three layers of our proposal and their associated methods, covering technical background knowledge and practice, local design conditions, and the system in operation.

This separation lays a conceptual grid over the manifold and distributed ecologies of normativity surrounding and pervading technical systems and serves to organize the analytical approaches we discuss in what follows. Our primary goal, here, is not to perform a full-fledged study of YouTube but to sketch a methodological toolbox and to propose an adaptation or recalibration of encircling to the domain of computational artifacts. However, taking a specific case as an example allows us to flesh out how one can apply and combine the various approaches, pointing to concrete materials and strategies for analysis as well as highlighting some of the difficulties and limitations that studying proprietary systems inevitably entails. The heterogeneous and multi-disciplinary nature of this endeavor means that parts of the analysis will vary in style and depth. Where possible, we focus on creating what could be called "analytical stubs", contained prototypes for how a larger analysis could look like and what one may be able to learn from it. For example, we briefly look at key papers from the history of recommender systems and publications coming out of YouTube through the lens of discourse analysis. For other methodological elements, however, we proceed more in the form of an evaluative literature review since setting up our own ethnographic study or data-driven analysis is beyond the scope of a single paper. While the particular weighing of elements is in part an effect of their varying degree of difficulty and cumbersomeness, it is also an effect of the exploratory process suggested by encircling itself: rather than follow a clear blueprint of methodological steps, it asks how we can "describe and analyse objects and terrains that are not directly visible for multiple reasons" (de Goede et al., 2019, p. 14) and the specific configurations of visibility and invisibility will depend in large parts on the case at hand.

Coming from three different disciplines — media studies, law, and computer science — that subscribe to wildly different epistemic cultures, we are not suggesting that our approach serves as some kind of meta-methodology that can integrate these disciplines into a singular perspective. Instead, the question of how technical objects integrate and operationalize nor-

mativity has created a productive space of encounter and exchange where different concepts and methods could be brought into conversation. When we argue that the combined presentation and discussion of these methods allows for drawing a more "holistic" picture of the various normative *forces* or *inputs* going into complex technical objects, we do not suggest that every research project has to combine our three layers and associated methods, but rather that there is value in realizing the many different dimensions that come into play. There is no single answer to the question this special issue asks, namely how to "explain machines" that resist explanation in fundamental ways (Burrell, 2016).

In his critical discussion of recommender systems as "traps", Seaver (2019) frames these systems not merely as socio-technical artifacts but argues that "epistemic and technical infrastructures come together to produce encompassing, hard-to-escape cultural worlds, at a moment when the richest companies in the world dedicate most of their resources to getting people hooked" (p. 423). Our approach seeks to contribute to our understanding of this world-making process, focusing on the production side and on the register of normativity, inquiring into the *materialization* of values across different spheres. Each of our layers adds to that process in specific ways and there is no singular "key" that unlocks the whole "secret". The specific combination of methods and disciplines we put forward, then, does not fuse into a singular perspective, but traces a fragmented progression across a number of dimensions.

While readers with different disciplinary backgrounds will have different levels of familiarity with the methodological strategies discussed in what follows, we believe that articulating them together can serve audiences in several practical and research contexts: for computer scientists and technical practitioners, a broader horizon with regard to normativity and how to examine it can alter the view of their own role as value-setting agents and potentially open pathways towards what Agre (1997b) called "critical technical practice", with "one foot planted in the craft work of design and the other foot planted in the reflexive work of critique" (p. 155). For legal practitioners and regulators, an understanding of values as spread over wider ecologies can help identify relevant points of contestation besides the algorithmic level. For researchers in the humanities and social sciences, the fields this paper most closely associates with, the different approaches discussed over the following sections can facilitate connecting with artificial systems that are notoriously difficult to study due to their complex technical character and corporate embedding. While several of these approaches require a certain level of technical skill, they point towards possibilities for transdisciplinary collaboration similar to the discussions that led us to write this paper.

3   Ambient Technical Knowledge

The question of how to handle elaborate technical subject matter is always pressing when approaching algorithmic systems from the perspective of the humanities and social sciences. One way to plot a path that is sensitive to technicity starts from the recognition that specific recommender systems do not exist in an intellectual vacuum. They stem from broader fields of knowledge and practice, and they draw on *archives* of techniques that have been developed, discussed, evaluated, and critiqued in computer science and adjacent disciplines. A new system never really starts from scratch, it draws on this background or *a priori* (Foucault, 1969) of established concepts, methods, and modes of problematization that are transmitted through textbooks, academic publications, and course curricula, as well as code libraries, Medium tutorials, and StackOverflow posts. Approaching technology as a field of knowledge and practice can be a way for scholars to gain technical understanding, to develop analytical categories for

the study of concrete cases, and to foreground the specificities of design and development activities. Dourish (2016) indeed calls for

> an alternative approach to algorithm studies which might put aside the question of what an algorithm is as a topic of conceptual study and instead adopt a strategy of seeking out and understanding algorithms as objects of professional practice for computer scientists, software engineers, and system developers. (p. 9)

Such an approach can take many different routes, including ethnographic and sociological investigation. Here, we focus on three directions that rely essentially on the analysis of widely available source material. *Critical reconstruction* constitutes a relatively "soft" entry point that allows us to identify and scrutinize basic ideas and principles that have structured the field over time. *Discourse analysis* extends this approach and focuses on key concepts, debates, and practices in the present. *Technical experimentation* proposes to submit available libraries for creating recommender systems to a form of "vivisection" that can sharpen our appreciation for the interaction between data and algorithm and serve as a prototype for the study of production systems.

3.1 Critical reconstruction

While a general historical anthropology of recommendation would provide an even more comprehensive starting point, a somewhat less expansive way to approach the field of recommender systems is what Agre (1997a) called "critical reconstruction", that is, the attempt to identify, examine, and interpret the "fundamental ideas and methods" of a field, in a way that "eases critical dialogue between technology and the humanities and social sciences". For us, this first means to historically trace and expound the various attempts to make the cultural gesture of recommendation a computable task. Recent examples of archeological or genealogical work in this direction (Mackenzie, 2017; Rieder, 2020) show how approaches sensitive to historical unfolding can draw on key publications and debates to reconstruct technical trajectories and narrow the gap between often forbidding technical detail and broader conceptual understanding. These texts are not simply taken as stepping stones toward the "state of the art", but as primary sources or materials that need to be submitted to critical analysis and interpretation.

In the case of recommender systems, one could start from Elaine Rich's (1983) *Users Are Individuals*, which presents Grundy, a personalized fiction literature recommender system that draws on explicit input in the form of self-descriptions and on implicit feedback based on the idea that "[o]ne of the simplest ways to derive information about a user is to look at the way he uses the system" (p. 205). Grundy was organized around a "user model" consisting of several facets, such as "interests" (e.g. "sports"), "politics" (e.g. "liberal"), or quantified psychological traits such as "tolerate-violence". These facets served as the features of a user-vector that was compared to the manually attributed book-vectors in the collection.[4] The system then began to suggest the book titles with the greatest overlap. As users provided feedback on recommendations, their profiles were updated accordingly. Jussi Karlgren's (1990) *An Algebra for Recommendations* is another early contribution and although it received relatively little attention, it proposed a forward-looking experiment involving a multi-user system and discussed implicit "grading" of documents through "behavioral criteria, like numbers of visits to it, access time,

---

4. Vectors, in computer science, are basically just valued (binary or numeric) lists where each entry encodes a variable or "feature". Calculating similarity or distance between two vectors underlies many approaches in information retrieval.

or other observable factors" (p. 3), anticipating the heavy reliance on empirical metrics of contemporary recommenders.

The system that is often seen as inaugurating "modern" recommender systems is GroupLens by Resnick et al. (1994). Unlike Grundy, which relied on competent staff for book classification, GroupLens was fully automated, in the sense that recommendations were made on the basis of comparing user behavior, leaving content characteristics to the side. This method, called "collaborative filtering", defines the basic setup of many systems to this day: "Collaborative filters help people make choices based on the opinions of other people. [...] The rating servers predict scores based on the heuristic that people who agreed in the past will probably agree again" (p. 175).

"Agreed in the past" is again established by comparing vectors, this time representing user profiles that register engagement with content items. If two profiles are found to be "similar", items "liked" by one user but unknown to the other can be recommended. This eliminates the need for editorial content classification and operationalizes recommendation as a purely behavioral function, as a *market* where users' actions determine the relative value of circulating items. This shift explains, at least in part, why contemporary recommender systems have great difficulties dealing with "unwanted" content, including misinformation or political extremism.

In 1997, Resnick co-edited a special issue of *Communications of the ACM* on recommender systems together with Hal Varian, a well-known microeconomists and Google's longtime chief economist. Their introduction not only summarizes different approaches to data collection and aggregation, but discusses social implications, business models, and the competitive advantages of having the largest user base possible. As is often the case in computer science, there is an "intuitive" justification and metaphorization of what is essentially a set of similarity calculations between users:

> In everyday life, we rely on recommendations from other people either by word of mouth, recommendation letters, movie and book reviews printed in newspapers, or general surveys such as Zagat's restaurant guides. Recommender systems assist and augment this natural social process. (Resnick & Varian, 1997, p. 56)

The identification of *problems* to be *solved* is indeed a locus of normative work in technical disciplines. The citations above indicate how technical publications formulate "problematizations" (Foucault, 1990) of the settings they intervene in and, in doing so, define their own specific ideas, techniques, standards, or "truth criteria". Key narratives, here, are often built around notions such as "information overflow" and "information need", which have dominated the field of information retrieval for decades (Rieder, 2020). But we also find elements that are more specific to recommender systems, e.g. that past behavior predicts future behavior or that people who have agreed in the past will again do so again, ideas that risk operationalizing a largely inertial view of people's preferences.

While a more stringent scientometric analysis is beyond the scope of our paper, identifying key publications is the backbone of "reconstructive" and hermeneutic historical work. This can include a survey of the dominant techniques in the field, which are broadly divided into content-based, collaborative, and hybrid methods. But as a field matures, it often diversifies and singles out problems that arise in specific application contexts. When looking at very large systems like the one at work in YouTube, issues like scalability (how to make "good" recommendations quickly as user base and content archive grow?) and sparsity (most users have only seen a tiny percentage of available videos) have received much attention. Techniques based on

matrix factorization and deep learning have been particularly successful at dealing with these problems.

Technical literature surprisingly often relies on psychological and sociological lay theories, often ignoring established knowledge from these disciplines. Evaluative practices that are organized around empirical measures have played a central role in keeping the need for deeper domain theorization at bay: success has most often been defined in terms of some form of congruence or overlap with actual user behavior, for example through lab experiments or in-situ measurement (Herlocker et al., 2004). How evaluation practices of recommender systems have evolved over time can thus be particularly revealing. In a recent paper, Seaver (2019) shows how "measures of success" have come to rely mostly on "captivation metrics" that are built on concepts like user retention and engagement rather than more contained definitions of "successful" prediction. Here, we find at least implicit affinity with concepts such as Samuelson's (1938) notion of "revealed preference", where behavior is taken as a direct expression of desire or volition, making behavioral signals unproblematic signs of user "satisfaction". Reconstructing these movements allows us to appreciate variation and contingency in technical fields and provides a structured way to work up to present day debates.

3.2 Discourse analysis

Clearly compatible with a historical perspective, (critical) discourse analysis (Foucault, 1969; Wodak & Meyer, 2016) can also be productively used to identify central tenets in contemporary recommender system thought and practice. Based on the idea that "language functions in constituting and transmitting knowledge, in organizing social institutions or in exercising power" (Wodak & Meyer, 2016, p. 7), discourses — understood as "relatively stable uses of language serving the organization and structuring of social life" (p. 6) — can be analyzed to understand how a group or field constructs and mobilizes a certain "worldview" that defines problems, concepts, norms, modes of admissible argumentation, and so forth. Taking sets of documents as material or data to be analyzed allows researchers to describe and analyze such discourses.

While we focus on academic material in our short dive into the field of recommender systems, industry publications and developer websites such as StackOverflow or Kaggle provide interesting windows into practitioner spaces, and local meetups could add an ethnographic dimension. But even analysis limited to academic literature already has to deal with huge quantities of published material. Anthologies and textbooks, such as the *Recommender Systems Handbook* by Ricci et al. (2015), which seek to package the "state of the art" into a single volume, are thus not only highly referenced and used heavily in teaching, but also helpful for analysts interested in examining a field from the outside. There are clearly many different aspects that one could take under scrutiny, but evaluation again stands out as a place where normative concerns are most explicitly formulated. One of the three chapters in the *Handbook* dealing with recommender system evaluation (Gunawardana & Shani, 2015), for example, specifies 14 properties or criteria that could be used as guiding values to design and optimize for: user preference, prediction accuracy, coverage, confidence, trust, novelty, serendipity, diversity, utility, risk, robustness, privacy, adaptivity, and scalability. The discussion of these properties yields a space of reasoning that structures normative concerns into addressable parameters and provides concepts and methods to talk about, justify, and evaluate design decisions, including potential trade-offs, for example between accuracy and diversity (Gunawardana & Shani, 2015, p. 280). Value articulations inform value practices, and metrics play a crucial role in turning broad ideals

into operationalized yardsticks.

While such lists of concerns can be useful, we also find more open-ended discussions in the literature, e.g. when Herlocker et al. (2004) debate why it is so difficult to evaluate recommender systems or when Jannach and Adomavicius (2016) survey and discuss different understandings of "purpose". The latter explicitly distinguish between "value for the consumer" and "value for the provider" (Figure 2), indicating that there is some level of contradiction or compromise between two "sides". Such lists can be used as conceptual frames for the analysis of concrete systems, including the purposes that play at best a subordinate role, e.g. the lack of effort to help users "understand the item space" on YouTube. In these areas, we also find particularly revealing instances of overlap and exchange between "cultural" and "technical" arguments, including a straight line linking technical features to product design and business planning.

Consumer's Viewpoint (i.e., value for the consumer)	Provider's Viewpoint (i.e., value for the provider)
Help users find objects that match their long-term preferences: Corresponds to the usual assumption that the main value of a recommender is to help users find relevant items in larger item sets. Recommendations could be limited to a subset of the items, e.g., new or trending ones.	*Change user behavior in desired directions*: Guide customers to other product categories, drive the demand from top-selling items to the long tail, leverage the persuasive potential of recommenders for up-selling purposes.
Actively notify consumers of relevant content: Proactively point users to new items through push notifications or newsletters, minimizing the user effort to check the site.	*Create additional demand*: Point users to other relevant items (e.g., accessories) to achieve cross-selling and advertisement effects.
Show alternatives: Recommend substitute products in the context of a reference item. A standard mechanism on e-commerce platforms.	*Increase (short term) business success*: Promote items with high margins or items that are in stock.
Show accessories: Recommend complementary products in the context of a reference item, e.g., with the goal of cross-selling. Also common in e-commerce settings.	*Enable item "discoverability"*: Increase the visibility of new items or niche products that would otherwise not be easily found through search or catalog browsing.
Help users explore or understand the item space: Help the user understand the space of options, possibly leading to higher choice confidence.	*Increase activity on the site*: Make users stay longer on the site, e.g., thereby increasing ad revenue.
Remind users of already known items: Provide users with reminders of repeated purchases of consumables. Or, present a list of recently viewed items that the user did not purchase so far; reminders then also serve as navigation shortcuts to a reduced choice set.	*Increase user engagement*: Increase customer loyalty and trust via a personalized service. Increase switching costs to other services as preferences are already known.
Improve decision making, e.g., in terms reduced decision time or higher choice satisfaction: E.g., provide the user with a limited choice set for an estimated purchase intent, provide explanations and interactive control.	*Provide a valuable add-on service*: Use the recommendation service as a differentiating factor from other competitors.
Establish group consensus: Provide recommendations that balance the interests of different group members.	*Learn more about the customers*: Utilize the collected preference information to better understand preferences and trends of the consumers. Provide mechanisms for users to explicitly state their preferences.
Help user explore: Provide a convenient way for users to browse the catalog without immediate shopping intent.	*Generate impression of dynamic, constantly updated site*: Dynamic recommendations contribute to the liveliness of the site. Updating the site with editorial content can be costly and less attractive for users.
Entertainment: Provide a satisfying emotional experience when visiting the site.	

Figure 2: Examples for purposes for recommender systems (from Jannach & Adomavicius, 2016)

These somewhat less visible domains of value articulation and arbitration are increasingly complemented by more explicit discussions of normative principles. Terms like "accountability", "transparency", "ethics", or "explainability" have become common within AI discourse and there is growing interest in what Morley et al. (2020) call "ethics tools", that is, "methods and tools available to help developers, engineers and designers of ML reflect on and apply 'ethics' " (p. 7). While it is hard to say how far into the field these efforts reach, they are showing a certain level of awareness for problems that were previously passed over and they are producing concrete techniques to identify and counteract bias and discrimination. At the same time, the problematizations that fuel this work are themselves circumscribed and limited to specific

sets of issues. This has led commentators like Powles & Nissenbaum (2018) to argue that "the preoccupation with narrow computational puzzles distracts us from the far more important issue of the colossal asymmetry between societal cost and private gain in the rollout of automated systems". Here, we are confronted with the distributed and layered character of normativity, making the question of which aspect or echelon to focus on a central part of value negotiation itself. Our goal is not to chart or adjudicate these discourses, but rather to highlight that technical fields like recommender systems mobilize normative reasoning on a number of different levels. Recent debates on AI ethics only capture the most visible parts of networks of normativity that reach deeply into all aspects of engineering practice. There is a need to investigate these less explicit layers, the conventions and "deep knowledge" that shape resulting technical systems in myriad ways, to gain a more comprehensive understanding of the full range of "ethical work" (Ziewitz, 2019, p. 713) that comes into play.

3.3   Technical experimentation

When dealing with technical artifacts, it is crucial to reserve space for technicity itself, to avoid a framing of technology as a mere epiphenomenon or transposition of cultural values. Here, we draw on Simondon's (1958) understanding of technicity as having meaning independent of use, that is, functional meaning that signifies through operation itself. If there is indeed something like properly "cultural" values, we can investigate how they structure professional practice, inform design decisions, and provide frameworks for evaluation. But "machine behavior" (Rahwan et al., 2019) — i.e. what a system actually *does* when embedded within a concrete application setting — does not follow teleologically from human intentionality. In large recommender systems, the sheer mass of participants and items has the effect that outcomes are best understood as metastable arrangements that depend on technical design, available content, and actual use practices, all evolving over time. The behavior of the system emerges from these different inputs and knowing how it works on the level of code is not enough to understand how it "reacts" when in use.

Researchers may not be able to gain access to a platform like YouTube, and almost certainly not in ways that permit hands-on experimentation, but there are open source libraries that — combined with public datasets — allow for forms of "vivisection" that can yield important insights. Projects like LensKit[5] or Facebook's DLRM (Deep Learning Recommendation Model)[6] may not be exact copies of real-world implementations, but they are part of the knowledge archives developers draw on and often used as models or starting points when creating new systems. For critical researchers, these libraries are prototypical examples of working systems, enabling a "hermeneutics of screwing around" (Ramsay, 2010) that can help develop intuitions for possibilities and limitations by playing with parameters, datasets, or optimization targets. Burell (2016), for example, tinkers productively with a simple spam filter based on a Support Vector Machine to support her argument that human and machine perception can differ in fundamental ways. LensKit for Python, "an open-source toolkit for building, researching, and learning about recommender systems" (Ekstrand, 2020), is particularly suited for similar research experiments, as it provides easy access to standard datasets and facilitates comparison between common algorithms. It implements many of the technical strategies and evaluation metrics discussed above, mapping that space of reasoning onto a modular technical artifact. While beyond the scope of this paper, technical experimentation can be a method

---

5.   See https://lenskit.org/

6.   See https://github.com/facebookresearch/dlrm

for investigating how broader normative concerns are operationalized and *materialized* into technical form.

4   Local Design Conditions

Ambient knowledge, practices, norms, and artifacts create an intellectual and material backdrop, but concrete projects involve design trajectories that are embedded in *local* circumstances. While task- and domain-specific knowledge circulates between the two levels and research fields are in constant contact with real-world environments, concrete application settings add many new commitments and concerns, often connected to the fact that recommender systems are run by for-profit businesses. Broad understanding of the technical field can inform investigations into actual systems by providing orientation and analytical categories, including a sensitivity for alternatives, but to step further into value construction as it unfolds in specific sites, another set of approaches is required. In this section, we discuss three methodological directions that case studies can follow. *Ethnographic investigation* is in many ways the "gold standard" for understanding local practices, even if platforms like YouTube are hardly open to academic scrutiny. *First-party documentation and "algorithmic gossip"* are valuable sources for cases where published material and informal community discussions are accessible. *Analyzing business models and legal context* allows us to situate systems in broader organizational environments that come with their own normative pressures, including incentive structures and legal conditions.

4.1   Ethnographic investigation

While workplace ethnographies have a long tradition and STS scholars have dedicated considerable attention to studying technical projects (e.g. Latour, 1992), research focusing on the practices and particularities of designing complex algorithmic systems are still rare. Seaver's (2017) study of a recommender company, which included interviews with about 40 people, is an exception that showcases the great potential of ethnographic methods in this field. What stands out in this work is how distributed, collaborative, and ultimately elusive "the algorithm" is, even in a company dedicated to technical production. There is no single person that has a complete grasp of the precise workings of the system. Recent work by Ziewitz (2019) targets a very different kind of technical field, search engine optimization, but pays particular attention to the way practitioners negotiate and reflect on their own, often morally ambiguous practices. We would argue that normativity is embedded in broader and often implicit considerations and processes, but we agree that "focusing on the work of 'being ethical' opens up a novel way of looking at the politics of algorithmic systems" (p. 725) in the sense that practitioners are not disconnected from or oblivious to material circumstances. While not the only "source" we can draw on, they are central informants for reconstructing the underlying constellations, including their material dimension.

But as Christin (2020) argues in her discussion of different ethnographic strategies for studying algorithmic systems, "the question of access remains crucial and complicated for ethnographers studying algorithms, especially on the construction side", since "[t]echnology companies and their engineering departments are deeply cautious and secretive, not only about ethnographers and academics but more broadly about all kinds of public discourse and reporting on their inner workings" (p. 913). We cannot count on companies like YouTube agreeing to anything close to the level of access Seaver (2017) was able to broker — although

the work of Klonick (2018, 2021) on Facebook's moderation practices and governance, which involved far-reaching access and peripheral participation, shows that openings do exist. But in most cases, former employees and whistleblowers are probably the closest one can get to genuine "internal" information.

What we learn from ethnographic work in related areas is that technical production comes with its own social complexities and, more often than not, bureaucracies and organizational idiosyncrasies. The different processes and negotiations that happen within design teams and the relationships with other organizational units are crucial to understanding how different factors — technical, economic, pragmatic, and properly "ethical" — are coming together. This includes broader ideological frames that are often left unexamined, for example the idea that more information is necessarily better or that every problem has a (technical) solution. While this points beyond what we are able to discuss in this paper, it should be clear that YouTube and similar companies are part of a larger "culture", whether one wants to call it "internet culture" or "californian ideology" (Barbrook & Cameron, 1996).

4.2 First-party documentation and "algorithmic gossip"

Even if "traditional" ethnographic methods such as interviews or participant observation are off the table for many of the most interesting cases, there are other sources to draw on that can provide insights into local circumstances, including first-party publications, official statements, and reports, but also documents and depositions shared during court cases. The people working on complex algorithmic systems are often (former) academics and many continue to publish. In fact, as Ahmed & Waheed (2020) have shown, large internet firms have come to dominate AI conferences, the central publication outlets in the field. In the case of YouTube[7], there are at least three papers (Davidson et al., 2010; Covington et al., 2016; Zhao et al., 2019) that provide substantial insight into the platform's recommendation system, how it evolved over time, and what kind of normative reasoning it draws on. While a deeper hermeneutics of these papers would merit another article, even a short discussion can demonstrate the potential of such an exercise for an analysis sensitive to value articulation.

First, the *overall framing* — or problematization — the system subscribes to is largely in line with the standard literature in understanding recommendations as helping users discover "relevant" (Davidson et al., 2010, p. 293) or "high-utility" (Zhao et al., 2019, p. 44) items. Interestingly, more philosophical musings, such as the idea that personalized recommendations address an "unarticulated want" (Davidson et al., 2010, p. 293), are no longer present in the later papers. There is overall a clear process of specialization in the sense that the initial view of the whole system, including interface aspects, yields to a focus on specific sub-problems and optimization puzzles. Second, this process of narrowing and refinement continues on the level of chosen *objectives*. While "freshness", i.e. the tendency to recommend recently uploaded videos, still features prominently in the 2016 paper, "diversity" disappears. But rather than a shift, this is more of a doubling down on the dynamic character of recommendations, which are no longer calculated in regular intervals, but when a user loads a page, increasing recency and making it possible to introduce "churn" in the sense that recommendations change when a page is reloaded (Covington et al., 2016, p. 6). Supporting viral spread becomes an explicit objective in this context. In the 2019 paper, the authors introduce the important distinction between "user engagement objectives" (e.g. clicking on video) and "user satisfaction objectives" (e.g. liking a video), which indirectly acknowledges the tensions between the consumer and provider

---

7.   A similar reading could be done on Facebook's work on recommender systems (Mudigere et al., 2021).

viewpoints discussed above. A more complete analysis could indeed draw on texts like the *Recommender Handbook* to trace which desirable properties are *not* explicitly articulated. Third, similar kinds of normative structuring appear in the discussion of the *user signals* to take into account. The reference to "privacy" disappears after 2010, giving way to the "data voraciousness" suggested and enabled by deep learning, where basically any kind of user behavior can be added productively to the model. Hundreds of features are being taken into account this way and the initial idea that users should "understand why a video was recommended to them" (Davidson et al., 2010, p. 294) is dropped. Fourth, the question of signal choice is particularly salient when it comes to the *evaluation metrics* used to assess and optimize the system, and one can again observe an evolution in the target measures. A/B testing on the live website is mentioned in all publications, but the 2016 paper now emphasizes watch time over click-through rate to combat "clickbait" videos (Covington et al., 2016, p. 5). The 2019 split between engagement and satisfaction, both captured via specific metrics, confirms the realization that a unique focus on the former may have detrimental impact for the site. The attention given to measuring and mitigating "representation bias", i.e. the tendency for highly recommended videos to accumulate more views and thus even more recommendations, again shows how the addressed issues are becoming more fine-grained (Zhao et al., 2019, p. 43). Here, we also notice the increased willingness to engineer "fixes" or "corrections" into the system, often driven by more nuanced modes of evaluation, resulting in a more complex architecture where the coordination of several independent modules generates the desired behavior. Fifth, these different considerations are constantly in conversation with the *technological choices* that translate desiderata into achievable features of the system. YouTube's move from pure collaborative filtering to matrix factorization and further to deep learning is first and foremost a testament to the immense importance that scalability, speed, and processing time have for the platform. The split into a less signal- or data-intensive candidate-generation stage, where millions of possible videos are reduced to sets of hundreds, and a much richer ranking stage, where many more signals are taken into account for a personalized ranking of these candidates, becomes the backbone of a system that prioritizes recency and reactivity.[8]

Taken together, these five points show a tightly connected problem space where engineers are not only "improving" recommendations, but developing an increasingly sophisticated and specialized understanding of a modular system that depends on their own technical work as well as on the behavior of very large quantities of users. Discussions of watch time, recency, virality, and so forth point to ideas about quality and desirability that are hardly at odds with economic success, even if the identification of risk and pitfalls (e.g. clickbait) leads to more nuanced apprehensions. Normative choices are operationalized and spread over a growing number of metrics that are discussed and balanced as "trade-offs". The cost of this specialization and modularization may well be reduced comprehension of the overall workings — and thus control — of the system, for both users and the engineers themselves.

A comprehensive analysis of first party statements would include not only academic publications but also official policy documents, technical documentation, media interviews, or financial reports as material to be analyzed. In line with the logic of "encircling", special attention should be paid to what is missing from these documents, how secrecy is operationalized and possibly changes in scope over time. For example, YouTube used to provide some basic insight into the main parameters of their recommendation engine (Google, n.a. a), before re-

---

8. We have argued elsewhere (Rieder et al., 2021) that the technical capacities that become visible here have an important political economy dimension and may play a crucial role in the continuing trend toward monopolization in the tech sector.

moving the page in 2015 and falling silent on the topic. While company statements need to be read with critical distance, they can reveal characteristics of working systems — or at least public justifications for design decisions. Facebook, for example, released a set of "recommendation guidelines" in 2020 (Fabeook, n.a.), which single out five types of content that "are allowed on our platforms, but that may not be eligible for recommendations". YouTube (2019a) follows a similar logic, even if policies are less detailed when it comes to defining "problematic" content. Here we learn that while the company has long deleted videos or channels that violate content policies, recommendation has become an important lever for acting on "content that comes right up to the line" (Youtube, 2019b). This includes limiting "recommendations of borderline content and harmful misinformation, such as videos promoting a phony miracle cure for a serious illness, or claiming the earth is flat". While these policies and their consequences are communicated in (strategically) vague language ("watch time that this type of content gets from recommendations has dropped by over 50% in the U.S"), they indicate a general shift towards explicit editorial intervention in systems that used to take pride in pristine automaticity. Here, we see how companies (sometimes) revise their value priorities in response to rising external pressure, which flags yet another area for future research, namely the question how far governmental scrutiny and critical commentary from civil society affect business practices and systems design.

Beyond first-party statements, we find another stream of material to study in absence of broad access, which comes from industry observers, consultants, and, most importantly, from participants that are directly affected and develop their own "folk theories of algorithmic recommendations" (Siles et al., 2020). In the case of YouTube, there are websites like Search Engine Journal[9] that aggregate and document the distributed knowledge production happening in the industry; analytics and marketing companies publish guides on how to use the platform's algorithms for growing one's audience (e.g. Cooper, 2021); and creators themselves continuously accumulate and share both deep practical knowledge of how recommender systems function and tactics for bending them to their own advantage. While this "algorithmic gossip" — "collaborative and directive processes used to formulate and sustain algorithmic expertise" (Bishop, 2019, p. 2589) — presents interpretative challenges, it can help generate singular insights into mechanisms and their effects, to the point of functioning "as a tool for exposing platform discrimination and bias" (Bishop, 2019, p. 2603). In her work on YouTube vloggers, Bishop not only shows how creators perceive a not-so-subtle preference for beauty content in YouTube's recommender systems, but also highlights the severe lack of communication and accountability when it comes to explaining how these systems work. While not the emphasis of our paper, this points towards the important question of how design choices and other corporate practices affect the different user groups gathering around algorithmically mediated platforms.

4.3   Business practices and legal context

A third approach to understanding a concrete recommender system in terms of value commitments and effects centers on the question of what or who it is working for. In the context of for-profit companies, this involves an analysis of the organization that designs and maintains the system, in particular its business model and the incentive structures that model generates (van Dijck et al., 2018). We certainly accept that the "bottom line" is not the only motivation for a firm and there are many ways to drive and develop sources of revenue; YouTube's boardrooms and legal offices are indubitably ripe with debate about how to react to external

---

9.   See https://www.searchenginejournal.com/

demands, for example. But it should be clear that private companies are mainly driven by a profit motive that exercises constant pressure on all design decisions. For YouTube, this means starting with the main source of revenue, advertising, which amounted to $15B in 2019, when Google broke out numbers for their video-sharing site for the first time (Statt, 2020). On the one hand, advertising pushes towards a logic of "more" — more viewers, watch time, engagement, etc. — and the revelation that 70% of traffic is steered by AI-based recommendations (Solsman, 2018) emphasizes the tight connection between these systems and economic success. The move toward "captivation metrics" (Seaver, 2019) that seek to tease users into staying on site is thus not surprising. On the other hand, recent years have seen multiple moments of public outrage and flashes in what has come to be known as the "Adpocalypse", where advertisers retreated from YouTube, forcing the company to adapt both policy and technology (Caplan & Gillespie, 2020; Kumar, 2019). The much greater willingness to make editorial decisions and the turn to more nuanced evaluation metrics are two direct reactions to this more complicated situation. Ranking, recommendation, and how these systems are communicated are indeed part of a balancing act that has to cater to different constituencies, including users, creators, advertisers, and regulators:

> Intermediaries like YouTube must present themselves strategically to each of these audiences, carve out a role and a set of expectations that is acceptable to each and also serves their own financial interests, while resolving or at least eliding the contradictions between them. (Gillespie, 2010, p. 353)

A critical analysis of the contradictory forces at play and the dynamics they produce will again not be able to produce definitive answers but serves to create an analytical scaffolding that connects the various other aspects coming into play to the broad imperative to generate profit.

An important part of a more theoretically informed analysis, here, is the realization that corporate practice unfolds within negotiation structures that are shaped by many different kinds of law. Competition law has recently received much interest in both Europe and the US, as commentators have argued for new yardsticks beyond effects on consumer prices (Khan, 2018). Labor relations are crucial for managing a workforce, in particular the army of low-paid moderators that sift through content at social media companies. Rules for data collection and ownership directly affect what signals a recommender system can draw on. Transparency requirements may force a company like YouTube not only to divulge specifications, but to change the way the system works in order to make it "communicable". Telecom regulation on issues related to net neutrality may affect bandwidth prices and the much-coveted speed at which recommendations can be served. These are just some examples from a panoply of rights and obligations that circumscribe companies' capacities to negotiate their participation within a market. When we ask what a system is "working for", we must thus add the question what it is "working with", that is, which background conditions shape its space of action. This includes a temporal component, as platforms like YouTube have been able to secure their position over time and under changing conditions, for example, drawing on personal data that is no longer legally available or issuing patents covering core components of their technical systems into the future.

Authors like Cohen (2019) have analyzed these complicated relationships between legal institutions and information technology in great depth, but even a more limited exploration can point toward instances where legislation has affected recommender systems very directly. Section 230 of the U.S. DMCA, in effect since 1996, and similar "safe harbor provisions" in other

legislations relieve "interactive computer services" from being liable for what users are sharing or posting on their services and explicitly allow for much leeway in content moderation, which recommendation is an increasingly important part of. More restrictive legal provisions, however, put limits on what data is collected and how it is made actionable. The U.S. Children's Online Privacy Protection Act (COPPA), for example, has led YouTube to introduce its "made for kids" feature, which excludes channels and videos targeting younger audiences from a number of features that rely on data collection. Recommendation is affected in the sense that videos made for kids are "more likely" to appear next to other such videos (Google, n.a. b). This means that once within the "kids sphere" users are no longer pulled into broader YouTube through recommendations. Here, the platform is effectively "deputized" by a regulator to enforce action against unwanted user-content encounters.

5   Materialized Values

The final layer of our methodological toolkit returns to the idea that technical artifacts are the result of value-laden choices and practices that become prescriptive and performative when their forms and functions intervene in concrete application settings. Normativity is *materialized* into the artifact itself, which may have far-reaching consequences for individuals and societies, but also means that technical forms and functions can become primary material for empirical ethical inquiry. As Bratton (2015, p. 44) writes, the "political program" of platforms can be found "in machines directly", raising the question whether and how we can "read" a system's operational values from the ways it looks and behaves.

In the context of social media platforms, Hallinan et al. (2021) discuss values embedded in and enacted through technological systems as "infrastructural values" that "may be expressed through interfaces, algorithms, APIs, engagement metrics, and reputation systems" (p. 6), already pointing to a set of specific components one may want to single out for analysis. Concrete methodological pathways for beginning such a project can also connect back to the technical experimentations with recommender system libraries we have mentioned above, focusing on "real world" systems instead. Many of the things we can do with these libraries are again off-limits for commercial systems, but in this section, we suggest interface analysis and data-driven observation as two strategies applicable to cases like YouTube's large-scale recommender system.

5.1   Interface analysis

In recent years, the notion of "affordance" (Gibson, 1986) has come to play a central role in the conceptualization and analysis of how material properties of objects and environments provide and suggest possibilities for action, linking them to human behavior. While the term has been used in design discourse to discuss artifacts and design decisions in great detail, more recent takes have emphasized "higher-level" affordances, such as the broader social and communicative possibilities offered by social media services (Bucher & Helmond, 2018). Postigo (2016, p. 1), for example, uses the concept to inquire how "technological features designed into YouTube create a set of probable uses/meanings/practices for users while serving YouTube's business interests", connecting specific features and properties of the platform to its central goal of revenue generation. In practical terms, the analysis focuses on the interface and the various functionalities that allow for "distribution of video, advertising, communication between commentators and subscribers, subscriber recruitment and retention, and community

participation" (Postigo, 2016, p. 13), creating an "architecture of digital labor" (p. 9) that extracts value from user practices. Approaches like "discursive interface analysis" (Stanfill, 2015), which seeks to examine "the assumptions built into interfaces as the normative or 'correct' or path of least resistance" (p. 1060), or the "walkthrough method" (Light et al., 2018) have attempted to organize similar forms of analysis into more structured methodologies. In both cases, the goal is to scrutinize the interfaces of apps and websites to understand how norms are constructed and user behavior is guided and shaped.

While these approaches do not focus on AI components, they draw our attention to the crucial question of how a recommender system is linked to use practices in functional terms. For YouTube, this means first of all localizing instances of recommendation in the interface, which is less straightforward than it may seem. While sections on the "home" and "trending" pages are clearly marked, it is less easy to understand which part of the "up next" column and even the search function are actually personalized to the user. We may also want to look at the leeway users have to configure, manage, or even circumvent the system. In line with previous observations, we notice that YouTube has hardly any way for the user to adapt the interface or tweak recommendations. For example, it is currently not possible to replace the heavily processed "Home" page with the "Subscriptions" page as default and giving feedback on recommendations is limited to a "not interested" button for individual videos, which can hardly be seen as a comprehensive and transparent form of steering the system. Note that affordances concerning interface components are relevant as they directly relate to users' actions on the system, and then to the production/transformation of value in different forms depending on the various stakeholders coming together around the platform (e.g. video creators, advertisers, etc.). These outer layers can and should be included in a more general study of (behavioral) affordances enabled by the system.

5.2   Data-driven observation

The second direction for studying materialized values draws on the idea that even proprietary systems are at least somewhat "observable" (Rieder & Hofmann, 2020) if their outputs can be scrutinized. Such "scraping audits" (Sandvig et al., 2014) can yield insights into the workings of AI components, even if a number of caveats apply. Other than the complexity of the underlying technical architecture, which may include many modules working together, and possible attempts to thwart data collection, the difficulty to "nail down" the mechanisms at work comes from the already mentioned recognition that "[m]achine behavior [...] cannot be fully understood without the integrated study of algorithms and the social environments in which algorithms operate" (Rahwan et al., 2018, p. 477). The behavior of YouTube's recommender system depends not only on purpose-driven and value-laden engineering feats but also on the practices of billions of users that upload, watch, like, comment, and so forth. How the system acts, what it recommends, how it evolves over time, and so forth, are *emergent* properties, especially when machine learning techniques are used. While the company's employees, crucially, define optimization targets and reward functions, actual outputs are not fully determined by these decisions, making value-focused analysis less straightforward than in the case of interface analysis. The presence of radical political content in video recommendations, a perpetual point of contention (Ingram, 2020), may variably be read as an indicator of YouTube's ideological preferences or as an effect of an editorial laissez-faire approach combined with the higher engagement propensity of divisive content. The question, then, is *whose* value commitments are at work here, a question that is exceedingly difficult to answer.

But if we approach the question from the perspective of performativity, we can decide to leave distributed agency entangled and focus on the recommender system itself as normative agent, examining its behavior in terms of normativity and decision-making. What is actually being recommended? Can we observe trends, for example, a preference for popular content or a tendency to drive users toward niches? How do recommendations evolve over time? A number of recent studies (e.g. Ribeiro et al., 2019; Alfano et al., 2020; Matamoros-Fernández et al., 2021) have analyzed recommended videos in this way, generally finding evidence in support of the radicalization hypothesis. Rieder et al. (2018), for example, focused on YouTube's search function and found oscillations between two ranking "modes", one prioritizing recency, where often divisive YouTube-native content dominates, and one that gives prominence to less polarized perspectives and more traditional media voices. While these forms of analysis come with their own caveats and are often difficult to implement, they allow us to investigate normativity as performativity, that is, as "values in action" rather than idealized input into design decisions. At the same time, we can conjecture that YouTube's engineers are well aware of how their system broadly behaves, which means that there is intentionality behind what we can observe.

6   Conclusion

This paper set out to assemble a methodological toolkit and research program for the (empirical) study of value(s) as they inform and manifest in concrete technical systems. While the often dominant narrative opposing "ethics" to "economics" played a role throughout, we have tried to focus on less visible vectors of normativity, including pressures and "inertias" that tie to longer historical trajectories and material circumstances. We were looking to multiply ways to understand what *goes into* the design and behavior of AI components, always understood as parts of larger systems. This has prompted considerations of method and methodology. The former concerns practical questions about how to investigate complex and proprietary technical systems, for which we have proposed the overall logic and method of encircling, adapted from security studies. Here, flexibility and multi-modal strategies are crucial. The latter concerns a broader understanding of values and norms, brushing against questions of epistemology when discussing engineering traditions and framing performativity as materialized values when looking at the forms and functions of working systems. A factor like "scalability" may not map onto anything like a first principle from moral philosophy, but it may well affect the choice of the central algorithmic technique for a given project and, in extension, the specific set of winners and losers the system designates.

How can a revisited form of encircling serve the study of such systems? What is the added value with respect to other approaches that combine different methods? First, an iterative, flexible, and multipronged approach can adapt to conditions of secrecy and to the considerable variation across cases. Specific methodological components are combined and calibrated in response to research possibilities and available analytical material. Second, compared to the approaches common in technical fields, whether they are code- or model-oriented (e.g. studying the data and machine learning methods applied to construct an AI component), user-centered (e.g. in the context of interface design), or focused on "machine behavior" (Rahwan et al., 2019) as an emergent property, encircling, as we envision it, covers more ground, in particular when it comes to understanding normativity as contingent on the ambient and local conditions design processes are subjected to. Third, with regard to the humanities and social sciences, encircling makes room for the integration of these more technical approaches as part of a multidisciplinary setup that refrains from taking values as exclusively cultural or social constructs.

Existing methods, taken individually, have only a narrow scope. In order to set up a view that is able to account for the various ways normativity weighs on technical systems, we need to cover the blind spots of different methods/disciplines, yet avoid reductionist shortcuts. Encircling, as an epistemic attitude and methodological strategy, provides a basis suitable to this project.

Using YouTube as our primary example has had the advantage of connecting to well-known social concerns and of providing a number of concrete materials and circumstances to analyze. But the exercise has also shown the immense difficulty to submit such a large, complex, and secretive platform to a multi-level examination, where each step could easily be a research paper in its own right. The reason why we opted to assemble eight research approaches into a collage of sometimes little more than investigative stubs, however, lies in the need to connect different disciplinary perspectives more tightly if we want to understand these expansive techno-socio-economic constructs. Encircling fragments in a hope to tie them together serves, in a sense, as a "best bad option" for creating a holistic view.

With regard to YouTube, a number of preliminary observations emerge: while we cannot *explain* the company's recommender system in the sense of identifying a list of causal variables or factors, we can argue that it largely sits within the tradition of collaborative filtering, where recommendations emerge from a behavioral market that is optimizing for a limited number of "captivation metrics" (Seaver, 2019). One of the effects of engagement led optimization may well be the tendency to recommend controversial and radicalizing content, as empirical audits have shown, or even more subtle effects such as the "preference" for beauty content identified in Bishop's (2019) work. But while the collaborative filtering approach, historically, was designed to do away with expert input, the balancing act the company has to constantly perform, catering to a number of different constituencies at the same time (Gillespie, 2010), means that the "editorial" inevitably creeps back in. The separation between engagement and satisfaction metrics, the fight against "clickbait", and the efforts to reduce representation bias indicate a more nuanced and goal-driven approach to the design of the recommendation market. And the attempts to decrease the visibility of "borderline content" and "harmful misinformation", or to single out content for kids, indicate that the company is willing and able to make editorial interventions to appease advertisers and regulators. At the same time, there is clearly little to no willingness to involve end users directly in these adaptations. The potential goal of helping users "understand the item space" (Jannach & Adomavicius, 2016) is not visible in YouTube's interface, the initial desire to explain why a video is being recommended disappears from the later technical literature, and there is hardly any means to configure or influence the system from the outside. None of these observations about the recommendation system's behavior amount to "opening a black box", but, taken together, they trace a space of interpretation and critique that is ultimately more salient than a purely technical interpretation.

To reiterate, our framework is an attempt to systematize the study of (proprietary and secretive) artificial systems along multiple layers, ranging from the broad normative commitments structuring a technical field to the emergent behavioral properties of a system in use, by going beyond the boundary/scope of individual disciplines. Coming from three different disciplines — media studies, law, and computer science — we did not seek to fuse the epistemological specificities of our disciplines into a flattened methodological apparatus, but juxtaposited different approaches in a way that keeps their respective identities palpable while allowing for connections to emerge out of these heterogeneities. This effort pays tribute to the multiplicity and distributed character of normativity in human affairs, even for something as seemingly simple as the question of how a technical system comes to frame and operationalize the idea of a "good" recommendation. But our program has clear challenges: it is wildly eclectic, hard

to implement without great effort, and not easily suited to be covered within one single study. To fully realize what we propose would require a heterogeneous group of researchers to invest significant amounts of time, similar to Eriksson et al.'s (2019) four year investigation of Spotify. But we think that researchers working on a single aspect or dimension of the larger assemblage — e.g. "algorithmic fairness" — can nonetheless profit from a broader understanding of normativity as spread out far beyond the ethical puzzles that have come to dominate the field.

While our purpose was not to examine the burgeoning landscape of deontological approaches to the design and critique of systems integrating AI components, we believe that the wider understanding of normativity we have championed over these pages is a necessary precondition for formulations of prescriptive ethics that "work", in the sense that they are capable of accounting for the wide array of forces that lay claim to the ways concrete systems perform in their application domains. For recommender systems, it is increasingly clear that focusing exclusively on popularity and engagement yields problematic results, but the larger framing of recommendation as a problem-to-be-solved, under the auspices of potentially reductive metrics, may require more attention and alternative narratives to allow for more substantial answers. Similarly, the lack of interest in providing functions for exploration, orientation, and adaptation in recommender systems like YouTube's may be understood as a concession to economic interests or as an attempt to avoid confusion, but it can also be seen as a disinterest in *training* or *teaching* users to navigate more proactively in the masses of content available. As AI-fueled systems penetrate into everyday experiences, Krug's (2014) famous mantra "don't make me think" is no longer just a design principle, but a potentially far-reaching social prescription that has been heeded all too well. While each of the layers we have delineated, here, is concerned with a particular aspect of a larger analytical puzzle, it can also be seen in terms of the opportunities and material for thinking about alternative arrangements to what we have now. These alternatives depend on more than finding the "right" deontological values and spreading them; they require the creation of assemblages and ecologies of value(s) where practices and materialities more broadly are investigated in terms of normativity, that is, in terms of which futures, like recommendations, they promote or suppress.

References

Agre, P.E. (1997a). *Computation and Human Experience*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511571169

Agre, P.E. (1997b). Toward a Critical Technical Practice: Lessons Learned Trying to Reform AI. In G.C. Bowker, S.L. Star, W. Turner, & L. Gasser (Eds.), *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide* (pp. 131–157). London: Psychology Press.

Ahmed, N., & Wahed, M. (2020). The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research. *ArXiv*, 2010.15581. http://arxiv.org/abs/2010.15581

Airoldi, M., Beraldo, D., & Gandini, A. (2016). Follow the Algorithm: An Exploratory Investigation of Music on YouTube. *Poetics*, *57*, 1–13. https://doi.org/10.1016/j.poetic.2016.05.001

Alfano, M., Fard, A.E., Carter, J.A., Clutton, P., & Klein, C. (2020). Technologically Scaffolded Atypical Cognition: The Case of YouTube's Recommender System. *Synthese*, *199*. https://doi.org/10.1007/s11229-020-02724-x

Ananny, M., & Crawford, K. (2018). Seeing without Knowing: Limitations of the Transparency ideal and its Application to Algorithmic Accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Association for Computing Machinery (2017). Statement on Algorithmic Transparency and Accountability. *ACM*, 12 January. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

Association for Computing Machinery (2018). ACM Code of Ethics and Professional Conduct. *ACM*, 22 June. https://www.acm.org/code-of-ethics

Barbrook, R., & Cameron, A. (1996). The Californian Ideology. *Science as Culture*, *6*(1), 44–72. https://doi.org/10.1080/09505439609526455

Belle, V., & Papantonis, I. (2020). Principles and Practice of Explainable Machine Learning. *ArXiv*, 2009.11698. http://arxiv.org/abs/2009.11698

Bishop, S. (2019). Managing Visibility on YouTube through Algorithmic Gossip. *New Media & Society*, *21*(11-12), 2589–2606. https://doi.org/10.1177/1461444819854731

Bonini, T., & Gandini, A. (2019). "First Week Is Editorial, Second Week Is Algorithmic": Platform Gatekeepers and the Platformization of Music Curation. *Social Media + Society*, *5*(4), 205630511988000. https://doi.org/10.1177/2056305119880006

Bosma, E. (2019). Multi-sited Ethnography of Digital Security Technologies. In M. de Goede, E. Bosma, & P. Pallister-Wilkins (Eds.), *Secrecy and Methods in Security Research: A Guide to Qualitative Fieldwork* (pp. 193–212). London: Routledge. https://doi.org/10.4324/9780429398186

Bratton, B.H. (2015). *The Stack: On Software and Sovereignty*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262029575.001.0001

Bucher, T. (2018). *If... Then: Algorithmic Power and Politics*. New York, NY: Oxford University Press.

Bucher, T., & Helmond, A. (2018). The Affordances of Social Media Platforms. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The Sage Handbook of Social Media* (pp. 233–253). London: Sage. https://doi.org/10.4135/9781473984066.n14

Burrell, J. (2016). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society*, *3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Caplan, R., & Gillespie, T. (2020). Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. *Social Media + Society*, *6*(2), 1–13. https://doi.org/10.1177/2056305120936636

Christin, A. (2020). The Ethnographer and the Algorithm: Beyond the Black Box. *Theory and Society*, *49*(5–6), 897–918. https://doi.org/10.1007/s11186-020-09411-3

Cohen, J.E. (2019). *Between Truth and Power: The Legal Constructions of Informational Capitalism*. New York, NY: Oxford University Press. https://doi.org/10.1093/oso/9780190246693.001.0001

Cooper, P. (2021). How the YouTube Algorithm Works in 2023: The Complete Guide. *Hootsuite*, 21 June. https://blog.hootsuite.com/how-the-youtube-algorithm-works/

Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. In S. Sen & W. Geyer (Eds.), *RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 191–198). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2959100.2959190

Davidson, J., Livingston, B., Sampath, D., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., & Lambert, M. (2010). The YouTube Video Recommendation System. *RecSys '10: Proceedings of the Fourth ACM Conference on Recommender Systems* (pp. 293–296). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/1864708.1864770

de Goede, M., Bosma, E., & Pallister-Wilkins, P. (Eds.). (2019). *Secrecy and Methods in Security Research: A Guide to Qualitative Fieldwork*. London: Routledge. https://doi.org/10.4324/9780429398186

Diakopoulos, N. (2015). Algorithmic Accountability: Journalistic Investigation of Computational Power Structures. *Digital Journalism*, *3*(3), 398–415. https://doi.org/10.1080/21670811.2014.976411

European Union (2020). Digital Services Act. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN

Dourish, P. (2016). Algorithms and Their Others: Algorithmic Culture in Context. *Big Data & Society*, *3*(2), 1–11. https://doi.org/10.1177/2053951716665128

Ekstrand, M.D. (2020). LensKit for Python: Next-Generation Software for Recommender System Experiments. *CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, (pp. 2999–3006). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3340531.3412778

Eriksson, M., Fleischer, R., Johansson, A., Snickars, P., & Vonderau, P. (2019). *Spotify Teardown: Inside the Black Box of Streaming Music*. Boston, MA: The MIT Press. https://doi.org/10.7551/mitpress/10932.001.0001

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press.

European Commission (2019). Ethics guidelines for trustworthy AI. European Comission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Facebook (n.a.). What Are Recommendations on Facebook?. Facebook. https://www.facebook.com/help/1257205004624246

Foucault, M. (1969). *L'archéologie du savoir*. Paris: Gallimard.

Foucault, M. (1990). *The Use of Pleasure. Volume 2 of The History of Sexuality* (R. Hurley, Trans.). New York, NY: Vintage Books. (Original work published in 1984)

Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., & Roth, D. (2019). A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. *ArXiv*, 1802.04422. https://arxiv.org/abs/1802.04422

Gert, B. (2004). *Common Morality: Deciding What To Do*. New York, NY: Oxford University Press. https://doi.org/10.1093/0195173716.001.0001

Gibson, J.J. (1986). *The Ecological Approach to Visual Perception*. London: Psychology Press.

Gillespie, T. (2010). The Politics of 'Platforms.' *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gunawardana, A., & Shani, G. (2015). Evaluating Recommender Systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (2nd ed., pp. 265–308). New York, NY: Springer. https://doi.org/10.1007/978-1-4899-7637-6_8

Google (n.a.). Creator Discovery Handbook. Suggested Videos on the Watch Page. Google. https://web.archive.org/web/20150329041618/https://support.google.com/youtube/answer/6060859?hl=en&ref_topic=6046759

Google (n.a.). Frequently Asked Questions about "Made for Kids". https://support.google.com/youtube/answer/9684541?hl=en#zippy=%2Chow-will-recommendations-work-for-made-for-kids-or-not-made-for-kids-content-will-the-discovery-of-my-videos-be-affected

Hallinan, B., Scharlach, R., & Shifman, L. (2022). Beyond Neutrality: Conceptualizing Platform Values. *Communication Theory*, *32*(2), 201–222. https://doi.org/10.1093/ct/qtab008

Hämäläinen, N. (2016). *Descriptive Ethics*. New York, NY: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-58617-9

Herlocker, J.L., Konstan, J.A., Terveen, L.G., & Riedl, J.T. (2004). Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems*, *22*(1), 5–53. https://doi.org/10.1145/963770.963772

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/1881.001.0001

Institute of Electrical and Electronics Engineers (n.a.). IEEE Code of Ethics. *IEEE*. https://www.ieee.org/about/corporate/governance/p7-8.html

Ingram, M. (2020). The YouTube 'Radicalization Engine' Debate Continues. *Columbia Journalism Review*, 9 January. https://www.cjr.org/the_media_today/youtube-radicalization.php

Jannach, D., & Adomavicius, G. (2016). Recommendations with a Purpose. *RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 7–10). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2959100.2959186

Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Karlgren, J. (1990). *An Algebra for Recommendations* (Working Paper No 179.). Department of Computer and Systems Sciences. KTH Royal Institute of Technology and Stockholm University. http://www.lingvi.st/papers/karlgren-algebra-for-recommendations-1990.pdf

Khan, L. (2018). The New Brandeis Movement: America's Antimonopoly Debate. *Journal of European Competition Law & Practice*, *9*(3), 131–132. https://doi.org/10.1093/jeclap/lpy020

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-offs in the Fair Determination of Risk Scores. *ArXiv*, 1609.05807. http://arxiv.org/abs/1609.05807

Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review*, *131*(6), 1598–1670.

Klonick, K. (2021). Inside the Making of Facebook's Supreme Court. *The New Yorker*, 12 February. https://www.newyorker.com/tech/annals-of-technology/inside-the-making-of-facebooks-supreme-court

Knobel, C., & Bowker, G.C. (2011). Values In Design. *Communications of the ACM*, *54*(7), 26–28. https://doi.org/10.1145/1965724.1965735

Krug, S. (2014). *Don't Make Me Think, Revisited: A Common Sense Approach to Web Usability*. San Francisco, CA: New Riders.

Kumar, S. (2019). The Algorithmic Dance: YouTube's Adpocalypse and the Gatekeeping of Cultural Content on Digital Platforms. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1417

Latour, B. (1992). *Aramis ou L'amour des Techniques*. Paris: La Découverte.

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. New York, NY: Oxford University Press.

Li, H.O.-Y., Bailey, A., Huynh, D., & Chan, J. (2020). YouTube as a Source of Information on COVID-19: A Pandemic of Misinformation? *BMJ Global Health*, *5*(5), 1–6. https://doi.org/10.1136/bmjgh-2020-002604

Light, B., Burgess, J., & Duguay, S. (2018). The Walkthrough Method: An Approach to the Study of Apps. *New Media & Society*, *20*(3), 881–900. https://doi.org/10.1177/1461444816675438

Lynch, M. (2001). The Epistemology of Epistopics: Science and Technology Studies as an Emergent (Non)Discipline. *American Sociological Association, Science, Knowledge & Technology Section (ASA-SKAT) Newsletter*, *Fall*, 2–3. https://asaskat.com/newsletters/

Mackenzie, A. (2017). *Machine Learners: Archaeology of a Data Practice*. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/10302.001.0001

Marcus, G.E. (1995). Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography. *Annual Review of Anthropology*, *24*(1), 95–117. https://doi.org/10.1146/annurev.an.24.100195.000523

Matamoros-Fernández, A., Gray, J.E., Bartolo, L., Burgess, J., & Suzor, N. (2021). What's "Up Next"? Investigating Algorithmic Recommendations on YouTube Across Issues and

Over Time. *Media and Communication*, *9*(4), 234–249. https://doi.org/10.17645/mac.v9i4.4184

Milano, S., Taddeo, M., & Floridi, L. (2019). Recommender Systems and their Ethical Challenges. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3378581

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Mudigere, D., Hao, Y., Huang, J., Jia, Z., Tulloch, A., Sridharan, S., Liu, X., Ozdal, M., Nie, J., Park, J., Luo, L., Yang, J.A., Gao, L., Ivchenko, D., Basant, A., Hu, Y., Yang, J., Ardestani, E.K., Wang, X., ... Rao, V. (2021). High-performance, Distributed Training of Large-scale Deep Learning Recommendation Models. *ArXiv*, 2104.05158. http://arxiv.org/abs/2104.05158

Nissenbaum, H. (1998). Values in the Design of Computer Systems. *Computers in Society*, *March*, 38–39.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Crown.

Pasquale, F. (2015). *The Black box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Pasquinelli, M. (2019). How a Machine Learns and Fails: A Grammar of Error for Artificial Intelligence. *Spheres*, *5*, 1–17.

Postigo, H. (2016). The Socio-Technical Architecture of Digital Labor: Converting Play into YouTube Money. *New Media & Society*, *18*(2), 332–349. https://doi.org/10.1177/1461444814541527

Powles, J., & Nissenbaum, H. (2018). The Seductive Diversion of 'Solving' Bias in Artificial Intelligence. *OneZero*, 7 December. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O., Jennings, N.R., Kamar, E., Kloumann, I.M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D.C., Pentland, A.'S.', Roberts, M.E., Shariff, A., Tenenbaum, J.B., Wellman, M. (2019). Machine Behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Ramsay, S. (2014). The Hermeneutics of Screwing Around. In K. Kee (Ed.), *Pastplay: Teaching and Learning History with Technology* (pp. 111–120). Ann Arbor, MI: University of Michigan Press. https://doi.org/10.2307/j.ctv65swro

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *CSCW '94: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work* (pp. 175–186). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/192844.192905

Resnick, P., & Varian, H.R. (1997). Recommender Systems. *Communications of the ACM*, *40*(3), 56–58. https://doi.org/10.1145/245108.245121

Ribeiro, M.H., Ottoni, R., West, R., Almeida, V.A.F., & Meira, W. (2019). Auditing Radicalization Pathways on YouTube. *ArXiv*, 1908.08313. http://arxiv.org/abs/1908.08313

Ricci, F., Rokach, L., & Shapira, B. (Eds.). (2015). *Recommender Systems Handbook* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/978-1-4899-7637-6

Rich, E. (1983). Users are Individuals: Individualizing User Models. *International Journal of Man-Machine Studies*, *18*, 199–214. https://doi.org/10.1016/S0020-7373(83)80007-8

Rieder, B. (2020). *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam: Amsterdam University Press. https://doi.org/10.5117/9789462986190

Rieder, B., & Hofmann, J. (2020). Towards Platform Observability. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1535

Rieder, B., Matamoros-Fernández, A., & Coromina, Ò. (2018). From Ranking Algorithms to 'Ranking Cultures': Investigating the Modulation of Visibility in YouTube Search Results. *Convergence*, *24*(1), 50–68. https://doi.org/10.1177/1354856517736982

Rieder, B., Sileno, G., & Gordon, G. (2021). A New AI Lexicon: Monopolization. Concentrated Power and Economic Embeddings in ML & AI. AI Now Institute. 1 October. https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-monopolization-c43f13698 1ab

Samuelson, P.A. (1938). A Note on the Pure Theory of Consumer's Behaviour. *Economica*, *5*(17), 61–71. https://doi.org/10.2307/2548836

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. Paper presented to "Data and Discrimination: Converting Critical Concerns into Productive Inquiry", a preconference at the 64th Annual Meeting of the International Communication Association, 22 May, Seattle, WA. https://social.cs.uiuc.edu/papers/pdfs/ICA2014-Sandvig.pdf

Seaver, N. (2017). Algorithms as Culture: Some Tactics for the Ethnography of Algorithmic Systems. *Big Data & Society*, *4*(2), 1–12. https://doi.org/10.1177/2053951717738104

Seaver, N. (2019). Captivating Algorithms: Recommender Systems as Traps. *Journal of Material Culture*, *24*(4), 421–436. https://doi.org/10.1177/1359183518820366

Siles, I., Segura-Castillo, A., Solís, R., & Sancho, M. (2020). Folk Theories of Algorithmic Recommendations on Spotify: Enacting Data Assemblages in the Global South. *Big Data & Society*, *7*(1), 1–15. https://doi.org/10.1177/2053951720923377

Simondon, G. (1958). *Du mode d'existence des objets techniques*. Paris: Aubier.

Solsman, J.E. (2018). CES 2018: YouTube's AI recommendations drive 70 percent of viewing—CNET. *CNET*, 10 January. https://www.cnet.com/news/youtube-ces-2018-neal-mohan/

Stanfill, M. (2015). The Interface as Discourse: The Production of Norms Through Web Design. *New Media & Society*, *17*(7), 1059–1074. https://doi.org/10.1177/1461444814520873

Statt, N. (2020). YouTube is a $15 billion-a-year business, Google reveals for the first time. *The Verge*, 3 February. https://www.theverge.com/2020/2/3/21121207/youtube-google-alphabet-earnings-revenue-first-time-reveal-q4-2019

Straube, T. (2019). The Black Box and its Dis/Contents: Complications in Algorithmic Devices Research. In M. de Goede, E. Bosma, & P. Pallister-Wilkins (Eds.), *Secrecy and Methods in Security Research: A Guide to Qualitative Fieldwork* (pp. 175–192). London: Routledge. https://doi.org/10.4324/9780429398186

van Dijck, J., Poell, T., & de Waal, M. (2018). *The Platform Society: Public Values in a Connective World*. New York, NY: Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

Van Veeren, E. (2018). Invisibility. In R. Bleiker (Ed.), *Visual Global Politics* (pp. 196–200). London: Routledge. https://doi.org/10.4324/9781315856506-29

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121–136.

Wodak, R., & Meyer, M. (Eds.). (2016). *Methods of Critical Discourse Studies* (3rd ed.). London: SAGE.

Yesilada, M., & Lewandowsky, S. (2022). Systematic Review: YouTube Recommendations and Problematic Content. *Internet Policy Review*, *11*(1), 1–22. https://doi.org/10.14763/2022.1.1652

YouTube (2019a). Continuing our Work to Improve Recommendations on YouTube. *YouTube Official Blog*, 25 January. https://blog.youtube/news-and-events/continuing-our-work-to-improve

YouTube (2019b). Our Ongoing Work to Tackle Hate. *YouTube Official Blog*, 5 June. https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/

Zhao, Z., Hong, L., Wei, L., Chen, J., Nath, A., Andrews, S., Kumthekar, A., Sathiamoorthy, M., Yi, X., & Chi, E. (2019). Recommending What Video to Watch Next: A Multitask Ranking System. *RecSys '19: Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 43–51). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3298689.3346997

Ziewitz, M. (2019). Rethinking Gaming: The Ethical Work of Optimization in Web Search Engines. *Social Studies of Science*, *49*(5), 707–731. https://doi.org/10.1177/0306312719865607

Zuboff, S. (2018). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York, NY: PublicAffairs.

Bernhard Rieder – Department of Media Studies, University of Amsterdam (The Netherlands)

 https://orcid.org/0000-0002-2404-9277

 b.rieder@uva.nl; https://www.uva.nl/en/profile/r/i/b.rieder/b.rieder.html

Bernhard Rieder is Associate Professor of New Media and Digital Culture at the University of Amsterdam and a collaborator with the Digital Methods Initiative. His research focuses on the history, theory, and politics of software and, in particular, on the role algorithms play in the production of knowledge and culture. This work includes the development, application, and analysis of computational research methods and the investigation of political and economic challenges posed by large online platforms.

Geoff Gordon – Asser Institute (The Netherlands)

 https://orcid.org/0000-0002-7067-1964

 https://www.asser.nl/ihcl-platform/about-ihcl-platform/who-is-who/GeoffGordon

Geoff Gordon is a Senior Researcher in International Law at the Asser Institute in The Hague. Researching governance issues at the interface of technology, security and economy, lately focusing on quantum information technologies, He has a background in litigation and an interdisciplinary academic background ranging across theory and practice with respect to global governance. His work is motivated by a critical concern for international practices in the public interest.

Giovanni Sileno – Informatics Institute, University of Amsterdam (The Netherlands)

 https://orcid.org/0000-0001-5155-9021

 https://www.uva.nl/en/profile/s/i/g.sileno/g.sileno.html

Giovanni Sileno is Assistant Professor at the Socially Intelligent Artificial Systems research group at the University of Amsterdam, member of the Civic AI Lab. With a background in electronic engineering, a PhD in AI & Law, and postdoc studies in cognitive systems and data-sharing infrastructures, he has been working in various fields related to AI and Computer Science research, such as computational legal theory, agent-oriented programming, cognitive modeling, computational policy design and operationalization.