Engineering Trustworthiness in the Online Environment

Hugh Desmond

Forthcoming in: Alfano, M., Collins, D. (Eds.). *The Moral Psychology of Trust*. Rowman and Littlefield

Abstract

Algorithm engineering is often portrayed as a new 21st century return of manipulative social engineering. Yet algorithms are necessary tools for individuals to navigate online platforms. Algorithms are like a sensory apparatus through which we perceive online platforms: this is also why individuals can be subtly but pervasively manipulated by biased algorithms. How can we better understand the nature of algorithm engineering and its proper function? In this chapter I argue that algorithm engineering can be best conceptualized as a type of environmental engineering aimed at making the online environment more hospitable to human use, in particular by safeguarding the conditions that allow for trust.

Keywords: Algorithms – Trust –Engineering – Control – Social Networks – Big Data – Freedom of Speech

1. Introduction

Algorithms are tools that empower users navigate vast online databases and communicate with each other in ways previously impossible. However, they also provide powerful incentives for certain types of social interaction, and hence restrict our social, professional, and public lives in new ways. Algorithms increase users' agency in some ways, and restrict it in others. How should we balance these two aspects in our conceptualization of algorithms?

The positive, empowering dimension of algorithm engineering was especially in focus in the early 2010s. For instance, social media were expected to re-energize democracies by allowing citizens to actively participate in public discourse (e.g. Loader and Mercea 2011). Evidence-based policymakers were expected to be able to fine-tune policy based on the data about citizens' preferences and behaviors (see Giest 2017). Scientists were to be empowered by publishing directly online, bypassing traditional gatekeepers such as peer-reviewers or editors (Bartling and Friesike 2014).

Today, the pendulum of public attention seems to have clearly swung towards ways in which algorithms encroach on freedom and even create further injustice. They are affecting how bank managers evaluate the trustworthiness of a mortgage applicant (Desai and Kroll 2017), or how a social worker evaluates the claim of a welfare applicant (Gilman 2021). They are affecting the shape of public discourse, giving outsized visibility to communications with strong emotions such as outrage (Munn 2020; Carpenter et al. 2021). And the open science movement has given rise to novel forms of inequality within the scientific community (see Desmond forthcoming). It has become easy to imagine a distinctly dystopian scenario, where machine-learning algorithms decide the shape of public discourse, determining which citizens to elevate in the social hierarchy, and which to keep away from both opportunity and privilege.

These phenomena raise the question: what precisely is the function of algorithm engineering? Should it aid users to find information that they prefer? Or should it aid users to find information that they *should* prefer? In other words, should algorithms strive to be as neutral as possible with regards to the type of social outcomes they might engender (e.g. outrage, polarization, discrimination) – or should they actively seek to promote certain types of social outcomes (e.g. honesty, trust) rather than others?

In this chapter I will sketch the contours of a fundamental framework for thinking about this question. The proposal will be to introduce the concept of the "online environment" as an abstract representation of the ensemble of particular online platforms, and then to foreground *information overload* as a structural feature of how individuals interact with the online environment. Information overload poses not just insuperable cognitive challenges for the human mind, but also significant social and moral challenges: what information should users *trust*? The human need for trust and trustworthiness points to the proper function of algorithm engineering: helping to shape the online environment in order to make it more hospitable for users to evaluate the trustworthiness of various instances of online communication – whether scientific publications, tweets and status updates, or loan applications. In other words, algorithm engineering should aim to safeguard the conditions of trustworthiness.

At first blush, it may seem strange or sinister to speak of “engineering trustworthiness”. Engineering may seem to be about control, whereas trust seems closer to letting go. Yet trust can only flourish under certain restricted conditions: when one is confident that the other is competent for instance, or does not have an incentive to lie. Moreover, if the “culture” -- the shared incentives and norms regulating speech and action -- is not conducive to trust, individuals may have difficulty finding sufficient reason to trust the claims of strangers. So in the same way an ecosystems engineer can intervene on variables in the environment in order to safeguard the habitats that allow species to flourish, an algorithm engineer may intervene on the online environment in order to safeguard the conditions for trust.

In the following section, I further develop these remarks about the relation between trust and control. I then introduce the concept of the "online social environment” and show how it makes sense that this environment is something that needs to be engineered. This would allow the precise ways in which algorithms impact trust to be elucidated, and the proper function of algorithm engineering to be defined. The last section discusses the prospects for making algorithm engineering more transparent, democratic, or professional.

2. Trust, Prediction, and Control

Control and trust are clearly distinct from one another. Control paradigmatically describes our attitude towards technological objects. When I flip a switch to turn on the light, one could say that I “control” the light. Our controlling relationship with objects is also apparent in our language: we use a “remote controller” for a television set, and the cockpit of an airplance is filled with “controls”. By contrast trust paradigmatically describes our attitude towards certain people like family or friends. Unlike control, trust implies that we are somehow vulnerable or not entirely certain about how the other will act (Baier 1986; Hawley 2014). When we trust a friend or colleague, we rely on their competence and goodwill, but we do not seek to control their behavior: betrayal remains a possibility. In sum, one can say that placing trust in someone involves relinquishing some control.

However, trust and control overlap in the sense that both involve some degree of *prediction*. If B is trusted by A, then A can predict (more or less) how B will behave and be pretty certain that B will not behave in untoward ways. A child is entrusted with a chore in the expectation they will carry it out; a physician is entrusted with a diagnosis in the expectation that the diagnosis will be carried out to professional standards. The expectation might not be met – the prediction may turn out to be false – but this means that trust turned out to be misplaced. Conversely, if the uncertainty about how B will behave is large – i.e., B could exhibit a wide range of action, not all desirable – then trust is not justified.

Because engineering is an area of human activity where trust and control become intertwined, I want to rephrase these remarks about prediction, control, and trust in terms of an influential account of causation, namely the manipulability account of causation (Woodward 2003). On this account of causation, "X causes Y" means that, if X is intervened on, that some change in Y will be observed.

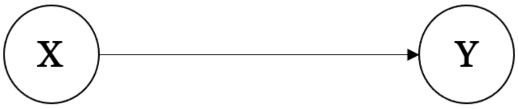

Figure 1: A dyadic causal relationship, or a dyadic trust relationship.

Strengths of this model of causal explanation include a relatively clear criterion to distinguish correlation from causation. For instance, it is incorrect to say that "the change in the barometer caused the storm", even though changes in the barometer reading precede changes in the weather pattern. The reason for this, according to manipulability accounts, is that if one were to intervene on the barometer reader, storm clouds would not form.

Now let us give Figure 1 a twist: let *X* be a speech act, and Y a behavior. For instance, A might ask B "can you pick me up from the airport" (speech act X) and then B might pick A up from the airport (behavior Y). The interesting question becomes: in what sense is this causal relation distinct from the relation between air pressure and a barometer?

Causal manipulability would entail that, if *X* is intervened on, and all other possible variables kept constant, then changes in *Y* would occur. Is that the case? Ordinarily not: the speech act does not "force" B to produce the behavior B, in the way that air pressure "forces" the barometer's needle to move. There is a crucial third variable, namely the intentional state of *B*: whether *B wants* to pick A up from the airport. Call this third variable *B*'s agency (β). Relationships of trust thus have structure as represented in Figure 2, where a speech act does not causally produce the behavior, but rather activates someone's agency, which they causally produces the behavior.

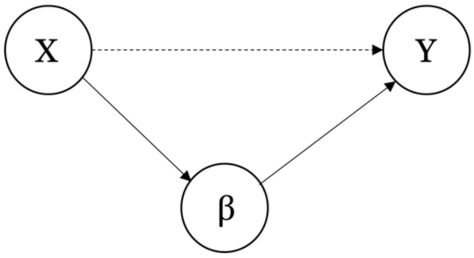

Figure 2: In a relationship of trust, the communicative act of A (X) does not control the response B (Y), but rather activates B's agency (β)

This is "ordinarily" the case, because some human relationships – those with great imbalances of power[1] – *do* come to approximate causal relationships. Let us take an extreme example for clarity's sake: *A* is a totalitarian dictator, and *B* is a chauffeur in *A*'s entourage. If *A* asks *B* to pick them up at the airport, *B* does not have much of a choice to refuse. The chauffeur may still choose to refuse and put to death as a punishment, so strictly speaking *B* still has some agency and of course *B*'s behavior is not literally controlled by the speech actions of *A*. However, the relationship between *A* and *B* is clearly not one of trust, and has come to approximate a brute causal relationship.

[1] Note that on this analysis, social power is intertwined causal power, and this calls to mind how "cause" and "power" were once used interchangeably (e.g., in the work of Hume or Hobbes). It goes beyond the scope of this paper to explore this any further.

In the following, we will be mainly focused on the perspective of the user of online platforms: the *receiver* of communicative acts. Questions about trust, from the perspective of *B,* concern aspects of *A's* agency: what are A's intentions or incentives? How should B decide whether to trust A's claim, and thus to endorse it? Figure 3 is the basic template we will be operating with.

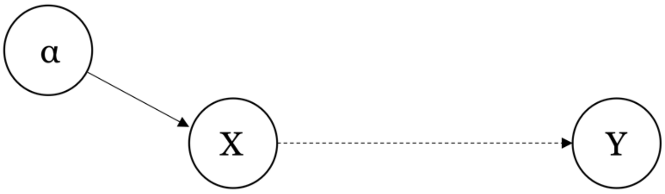

Figure 3: A's agency (α) is the crucial variable for B to evaluate the trustworthiness of X.

Further, on online platforms, a receiver *B* may need to evaluate *many* senders *A, A', A",* and so on. As B scrolls down their news feed, the interactions between B and A' or between B and A" will be fleeting, depersonalized, and superficial. Due to limited cognitive processing power, *B* will not be able to even fleetingly and superficially evaluate every message being posted. The receiver B will need prior heuristic decisions on which messages are *worthy of attention:* in other words, a selection needs to be made on claims that are even *potentially* endorseable. This means that the question of trust in the online environment not only turns on the evaluation of the content of the message, but also on the selection of which messages to give attention. The next section clarifies some these basic features in the online social environment.

3. The Online Social Environment

With the "online social environment" I mean the ensemble of communicative acts that are stored on websites : personal and blogging websites, social media websites, scientific repositories, institutional websites, and so on. It is the part of the internet where humans are directly interacting with other humans.

The online social environment could be parsed in several ways. One structural feature of this online social environment is that a lot of it contains *public communication.* In other work (Desmond forthcomingb) I show how social media promotes the public communication of private content. This is less of a challenge for algorithm engineers rather than users of social media, who must develop virtues in dealing with social media in a mature way.

The feature of the online social environment that will be foregrounded in this chapter is *information overload*. By “overload” I mean that the number of written communications that are published to various webpages – personal, institutional webpages, social media sites, scientific repositories – is so large that a single user cannot read more than a minuscule fraction of them. Information overload is a basic and unavoidable feature of the online social environment: it can be understood as a direct consequence of the main attraction of online communication, namely lowered costs and barriers to communication. , but information overload presents novel problems for decisions of trustworthiness.

Just how big is the information overload in the online social environment? If one were to measure it with the number of websites in existence: 1.9 billion in 2021 (Statista 2021). However, some individual websites present their users with information overload. In 2014, there were an average of 500 million 140-character messages posted to Twitter every day (Twitter 2014), or almost 6000 per second. Even the number of published scientific articles – many orders of magnitude smaller than the number of published tweets – presents information overload: over 3.1 million articles published in 2022, or about 8500 articles per day (see (Desmond 2021a) for calculation). Even just scientific communication – a miniscule fraction of all communications posted daily – vastly outstrips any one individual’s processing capacity. We often marvel at the vastness of the universe compared to the Earth: one could similarly marvel at the vastness of communications being generated by humanity compared to the tiny cognitive niche each individual occupies.

In an environment with information overload, two major challenges emerge: (1) evaluating whether a communication should be believed (or trusted), (2) evaluating whether communications are worthy of being evaluated for trustworthiness. In the online social environment, where users can come into contact with a vast number of unknown senders, it can be difficult to evaluate whether or not to believe their communication. However, the process of deliberating on whether to believe a communication or not itself requires cognitive investment. Users cannot evaluate the claims of every single communication being posted; moreover, many communications do not contain worthwhile information for an individual user. So in environments with information overload, users must rely on external processes to make a selection of what communications they should submit to further evaluation.

These external processes are *gatekeeping* processes. In a library – until the advent of the internet, the social environment most paradigmatically characterized by information overload – there were at least two gatekeeping levels: librarians who chose which books to buy and where to display them; editors and reviewers who chose what manuscripts to publish as books. In the online social environment, algorithms have become an additional (de facto) gatekeeper, "deciding" which communicative acts to rank higher on search results than others.

4. Engineering and the Environment

Both the environment and engineering paradigmatically refer to physical spaces and processes: the environment referring to the physical processes impinging on an organism, and engineering involving the construction of artifact. By contrast, the online social environment seems to be constructed of abstract entities: online communications. Algorithms then are constructed to sift through these abstract entities: what does algorithm engineering have in common with, for instance, civil engineering? In this section, we digress briefly on how the online social environment can be understood in continuity with biological environments, and how engineering in general refers to human attempts to shape their environments.

How to situate the online era in the long arc of human evolution? Today, in the 21st century, rival predators no longer pose a threat, and climate change notwithstanding, weather systems pose far less of a threat than they once did (Ritchie and Roser 2014). Microorganisms – bacteria, viruses, parasites – remain considerable threats, but again this is relative. If the severity of the threat is compared to our evolutionary past, then it would also seem much diminished if one looks at the impact of infectious infant mortality and maternal mortality during childbirth. Both have collapsed from the levels that characterized most of human evolution, from 50-100 maternal deaths per 1000 live births to about 0.1 in developed countries today (Chamberlain 2006) and from over 300 infant deaths in their first year of live per 1000 live births to 3 today (Mühlichen, Scholz, and Doblhammer 2015).

The ecological threats that dominated humans' evolutionary past may be much diminished, but this is not necessarily an argument for inexorable human progress (such as in Pinker 2018). My point is rather that the main source of ecological threat has shifted to the *social* environment. Our main sources of worry are not wolves or disease, but human beings. We worry about criminality, traffic accidents, social rejection, depression. Other humans remain the greatest source of opportunity for each other, but they are also increasingly each other's greatest threat.

With this generalization I do not wish to downplay the history of warfare. Rather, I wish to point to how the increased connectivity between individuals is leading to new, acute sources of distress. The challenges of hate and harassment were already mentioned, but to just add one concrete illustration of the generalization: consider how, between 2000 and 2015, the suicide rate for teenage girls doubled. Between 2000 and 2010 there was a relatively slow increase, with an accelerated increase occurring after 2010. Jean Twenge and colleagues (Twenge et al. 2018) have argued that social media has at least a large part to play in this increase: the smartphone was introduced in 2008, and this allowed social media could be accessed at any time and in any context. From that moment on, social media became intertwined with teenage life, and with it, the intense pressures of social approval and disapproval.

It remains unclear just how harmful these novel challenges are in the online environment. Some claim digital harm is overestimated while others argue it is probably still underestimated (see Twenge et al. 2020). Regardless of just how harmful it is, from a broader evolutionary perspective it is striking that changes in pixels on a screen could have such a large impact on the health of an individual. It illustrates how the most salient threats in the external environment no longer consist of pathogenic micro-organisms or weather systems, but increasingly of speech acts that provoke jealousy, hatred, rage, or depression. These dangers are the basic rationale for engineering trust – in the way we engineer houses to shelter us from the elements, so we should engineer algorithms to shelter users from abuse.[2]

Another plotline in human evolutionary history has been *niche construction*, which refers to how organisms can actively shape natural environments in order to

[2] The analysis in this chapter remains neutral on ongoing political discussions about whether speech should be regulated in the online environment. As will become clear later, there can be no serious question about *whether* speech should be regulated (certain speech acts, such as defamation or incitement to genocide, are illegal in most jurisdictions). The difficult questions – on which this chapter takes no stance – arise concerning how strict these regulations should be (i.e., how precisely does one define defamation or incitement).

alter selection pressures (Laland, Matthews, and Feldman 2016). Niche construction is a widespread response to change being the default state of natural environments (Desmond 2021b). Not all such change is favorable to an organism: abiotic factors (nutrients, temperature, humidity, acidity, etc.) may fluctuate, as may biotic factors (predators, competitors, potential mates, etc.). In response to such changes, humans have built shelter, weaved fabric to cover their bodies, or started fires: all these actions modify the surroundings (temperature, humidity) of the human organism as to make them more favorable. Humans are not in the least the only species to engage in niche construction (a classic example is the beaver building dams), but insofar as niche construction should be understood as an agential activity (Aaby and Desmond 2021), it is not surprising that human agents have been very active in shaping their immediate surroundings in favorable ways. This is apparent in humans' surprisingly large geographic extent and population size (Desmond and Ramsey in press). Humans have been so active in shaping (and exploiting) their surroundings that geologists are increasingly confident that a new geological epoch needs to be named after the human species, namely the Anthropocene (see e.g. Zalasiewicz, Williams, and Waters in press).

Engineering can be understood as a type of niche construction. In the popular imagination engineering is often understood in much more narrow way, as concerned with a science-based design and production of *machines* (Mitcham 2020). This narrow focus on machines is suggested by the very word of engineering ("engine"). However, it is not necessarily helpful in understanding the broad variety of types of engineering, both today and historically. A civil engineer designs buildings, bridges, and larger infrastructure projects, such as dams. An environmental engineer may analyze soil components and tackle ground pollution. Neither civil engineers nor environmental engineers deal with the design of machines or artifacts, but both engage in niche construction.

Historically, the construction of irrigation structures in ancient Mesopotamia and Egypt is often cast as the first appearance of engineering (Alexander 2020, 27). These structures consisted of dikes of earth and reeds, were labor-intensive, and would have required extensive planning. Their function was to buffer against variation in rainfall: the dike system allowed flood waters to be captured and distributed over a much larger area of land than would have otherwise been the case.

The deeper etymology of engineering is not engine, but the Latin *ingenium,* capacity or ability.[3] And in one of the first definitions of the engineering profession, the British Institution of Civil Engineers described its activities in 1828 as "the art of directing the great sources of power in nature for the use and convenience of man" (Mitcham 2020, 12). This points to how broadly engineering can be defined, as a fundamental human activity. For instance, another broad definition of engineering is that it is "the use of heuristics to cause the best change in a poorly understood situation within the available resources" (Koen 2003; cited in Mitcham 2020, 12). Such definitions may be overly broad, but they do acknowledge the great variety in types of engineering. The definitions also acknowledge that, while the profession of engineering today is a science-based activity, engineering predates science by millennia.

Both these earliest forms of engineering as well as some of the broader definitions of engineering illustrate how engineering, at its most fundamental, is an activity that aims at the *control* of certain aspects of the external environment. It does not aim at scientific understanding, nor does it aim at a religious acceptance of the world as it is. As a form of niche construction, engineering aims to buffer against changes in the environment, and reshape these changes in ways that are advantageous for humans.

This perspective on engineering involves redescribing the activity in environmental terms. For instance, the aeronautic engineer designing aircraft aims at allowing humans to take advantage and even *control* the lift that fast-moving air provides. The civil engineer designs infrastructure to take advantage and control certain variables in the natural environment. A house controls the ambient temperature and humidity for its inhabitants; a dam controls the flow of water through a river. Also a military engineer (historically the counterpart of the civil engineer) seeks to control the environment, for instance, by controlling projectile motion in various ways, whether the projectiles are launched from catapults, trebuchets, canons, guns, or tanks.

It is in this fundamental sense of engineering – designing and producing artifacts that allow humans to control the environment – that it makes more sense to speak about algorithm engineering. Like an irrigation engineer seeking to control the

[3] A yet deeper etymology traces it back to *gignere* (to beget, give birth, or to cause).

flow of water by designing canals and trenches, an algorithm engineers seeks to shape flows of information in the online social environment by designing algorithms. The complication for this analysis is that, for the algorithm engineer, the targeted environment is *constituted by* the users. Thus, by empowering a user to find the communications that are useful for them, simultaneously the algorithm will *restrict* the user from interacting with other communications.

This tension between empowering and restricting is present in many if not all forms of engineering. Consider the automobile. At the time of their introduction, automobiles clearly empower their users compared to the rivals of horses or trains: users could choose to travel greater daily distances and yet with more control over departure time as well as destination than would have been possible by train. Yet, the success of the automobile set in motion a chain of events that ultimate restricted citizens' choices in other ways. Most saliently, it promoted a "sprawl" in urban and suburban planning, which in turn led to commerce and public facilities being located non-walkable distances from each other and from housing (see Flink 1988). The automobile thus has come to indirectly dictate certain aspects of individuals' lifestyles.

The tension between empowering and restricting is a consequence of how engineering reshapes the environment: it reshapes the environment according to human needs (empowering aspect), but once reshaped, the environment provides new incentives for behavior that adapts to the environment (the restricting aspect). This general tension is thus inherited by algorithm engineering, but the empowering and restricting become more intertwined because the users constitute the relevant environment.

5. Trustworthiness in the Online Social Environment

In many analyses of trustworthiness (e.g. Hawley 2014), two conditions for trustworthiness are typically distinguished: the competence (or expertise) and the intention of the trustee. When we make decisions whether or not to trust a person, we evaluate evidence for their competence and intentions. For purposes here, we do not need to delve into the details of this epistemological reasoning; instead, what I would like to highlight is how our evaluations of trustworthiness depend on the judgment of others.[4]

5.1 The Social Embeddedness of Trustworthiness

First, we may evaluate competence of a person while lacking the relevant knowledge to do so independently. Instead, we may rely on indicators such as their job title, diplomas, degrees, and so on. This can be justified insofar as such indicators signify how others with relevant knowledge have judged the trustworthiness of that person. In evaluating the truthfulness of a claim, we may also further rely on indicators such as the venue where the claim was made, i.e., whether (for instance) the claim was made in an online forum where teenagers joke around with each other, or in a national newspaper with editorial review. If a claim is published in the latter venue, the receiver of the message may presume that the claim would have been vetted by the editorial team, and hence the receiver may come to rely on the judgment of those gatekeepers. In sum, the evaluation of a sender's competence (or the truth of the message) often in practice depends on how others evaluate the sender's competence (or the message's truth).

Not only our evaluations of competence, but even our evaluation of the *intentions* of others depend strongly on the social network we are embedded in. Speech acts are assumed to obey certain norms of honesty, as well as certain incentives for truthful speaking. For instance, even if we do not personally know whether the speaker is reliable, if they run into a crowded theater shouting "fire", it would not be naïve to presume they are telling the truth. After all, some instances of lying can be prosecuted under the law. False speech can cause great harm, and the category of free speech in many jurisdictions does not include the following:

- Defamation: making false allegations of immoral or criminal conduct with the purpose of harming a person's reputation.
- Incitement to hatred: making false statements about a group of persons with the purpose of increasing discrimination against that group.

[4] So not only can testimony be a reliable source of knowledge, but in order to evaluate whether a testimony is trustworthy, we in practise rely on yet further testimony.

- Incitement to genocide: making statements about a group of persons with the purpose of promoting extermination of that group.

In a culture where speech acts of defamation and incitement are strongly disincentivized, it is not naïve to assume that the average citizen will have at least some incentive to not commit those forms of false speech, and thus to adhere to certain minimal standards of honesty.

While certain minimal norms of honesty in virtue of a shared legislative context, norms of honesty can also vary considerably given the professional context. For instance, most professionals are expected to adhere to a code of conduct, which almost invariably includes some provisions regarding honesty and diligence. When a physician communicates a diagnosis, the patient can assume that the physician has taken all reasonable precautions to be able to establish an accurate diagnosis (Desmond and Dierickx 2021). Other norms of honesty hold in the media environment. Journalists who reporting on events are held to high standards of accuracy, at least according to codes of journalistic ethics (see e.g. SPJ 2014). By contrast, different standards of accuracy are permitted for polemic journalists in their columns or op-eds: there it is accepted that they exaggerate with the aim of provoking the reader.

Much more could be said about social norms and incentives regarding honesty: the lesson for the purposes of this chapter is that an agent's evaluations of the trustworthiness of another agent, whether in terms of competence (expertise) or honesty, is dependent on social norms of various kinds, and especially laws and professional norms. Evaluations of trustworthiness are social processes. Without being able to rely on the judgments of others, or on honesty norms, it would become much more difficult – and likely impossible in many contexts – for agents to evaluate the trustworthiness of a stranger. This reliance is only intensified in the online environment, but with the challenge that laws and professional norms are too weak to regulate trustworthiness of speech.

5.2 Trustworthiness Evaluation in the Online Environment

A relevant difference between online and offline social environments is the degree of interconnectedness between a vast number of senders and receivers.

Reaching a large number of agents with a single communicative act was already made possible by mass communication technologies (radio, TV) and arguably print (pamphlets). However, such forms of communication are highly centralized, with only a small number of senders. This means that a receiver need only evaluate the trustworthiness of the small number of senders (i.e., news anchors, radio hosts, journalists, editors, and so on).[5] By contrast, on platforms such as social media, all agents are potential senders and receivers, and the network structure is more decentralized.

One type of minimal value-ladenness of algorithm engineering follows from information overload. Search algorithms work by foregrounding some communications, and they aim to do this in such a way as to empower users, by helping them to find the communications they are interested in. However, by forgrounding some communications, algorithms necessarily background others. By determining what communications is potentially trustworthy, algorithms also determine what communications are not even worth the user's time. In this way, algorithms shape the online social environment and hence restrict users in other ways.

As a concrete illustration of this dynamic, consider some of the basic features of YouTube's recommendation algorithm – or at least, how the algorithm was operational state in 2016. According to public sources currently available, this recommendation algorithm largely evaluates videos according to *engagement* (i.e., how likely users will click on it). Videos that have more engagement are ranked higher than videos that have less engagement; the higher ranking leads to the high-engagement videos gaining more engagement, and so on. Engagement is prioritized not just for monetary reasons (i.e., to show users advertisements), but also because engagement is assumed to be a good measure for what users value. If a user clicks on a video, the assumption is that they value the content in that video (and so in this way, algorithms can tailor their recommendations based on a user's history).

[5] Conversely, only a small number of senders need to be manipulated in order to mislead a large number of receivers: propaganda.

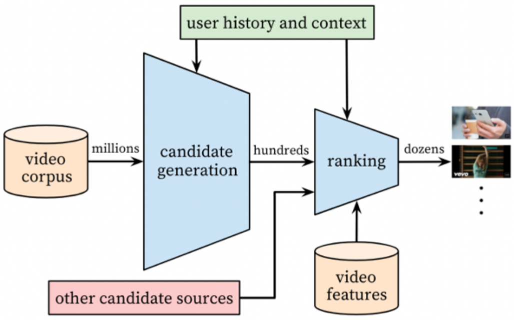

Figure 4: A visual representation of YouTube's recommendation algorithm. Reproduced from (Covington, Adams, and Sargin 2016)

However, despite attempting to design a relatively neutral tool that tracks user preferences, such algorithms end up creating novel incentives for certain types of speech act rather than others, and thus shaping discourse in ways that were not initially intended. In particular, this prioritization of engagement is also often cited as the fundamental reason why content that promotes moral outrage tends to be disproportionately present in users' feeds (Munn 2020; Alfano et al. 2021). In this way, by selecting what communications are worthy of user's attention, algorithms influence evaluations of trustworthiness. Some types of communications are never shown to users; other types are shown too often. This is the first important way in which algorithms shape patterns of trust and distrust in the online social environment.

A second consequence of information overload is that legal and professional norms are insufficient to regulate honesty in online environments. On a social media platform such as Twitter, with hundreds of millions of users, users have highly heterogenous backgrounds, and cannot be expected to collectively endorse any type of demanding professional or editorial norms of honesty.[6] Moreover, even if some norms of honesty could be expected of the hundreds of millions of users, they would not be enforceable. Even very minimal norms regarding defamation and incitement cannot be enforced: assuming they occur with some infrequent regularity on a platform such as Twitter, with 500 million tweets per day, allegations of defamation and incitement would quickly overwhelm courts of law.

In this way, algorithms must also take over some of the regulative work done by professional and legal speech norms. Here algorithms directly restrict users to

[6] Moreover, communications on social media often contain private content, and so often cannot be considered merely public or professional communications. In (Desmond forthcomingb) I analyze in more detail how the public and private become intertwined on social media.

prevent problematically dishonest speech acts from propagating, especially defamation or incitement. This is inevitably politically sensitive, because the question of what constitutes incitement is political: it turns on the question what is *justified discrimination* and what is *unjustified discrimination*.[7] Nonetheless, if very basic norms of honesty are to be upheld, the question is not *whether* speech must be regulated by algorithms, the question is *how*.

These remarks about speech regulation may seem more controversial than they are, because the early days of the internet were governed by a more anarchic-libertarian philosophy of unfettered free speech. Hence we should consider the contrasting case: could we let norms of honesty be spontaneously regulated by users? The first problem in doing so is that the assumption of honesty is what drives engagement with communications of others: if we know someone is lying, we will not believe their claims are are less likely to even pay attention. Hence, allowing norms of honesty to erode entirely would ultimately damage a platform. Second, considerations of justice require some top-down regulation of honesty. Defamatory and inciteful statements can cause genuine damage to their targets, and in subsequent court cases algorithm engineers (and corporations) could be found to bear some culpability if their algorithms allowed such statements to spread unchecked despite protestations.

6. Algorithm Engineering as Trust Engineering

If one were to generalize over the online social environment as a whole, promoting the conditions of trustworthiness would seem like a good candidate for the proper (and default) function of algorithm engineering. This means that algorithms need to be designed in such a way that more trustworthy communications are more visible than less trustworthy communications – and trustworthiness is not the same as popularity, or as engagement. The challenge for (good) algorithm engineering is thus to find ways to recognize trustworthy communication, whether in terms of competence or honesty.

Stipulating trustworthiness as the proper function of algorithm engineering does not imply that, in certain contexts, algorithm engineers can deviate from that function. For instance, on an entertainment website such as Netflix, it is defensible

[7] For instance, in many societies, there is widespread discrimination against people with a criminal record. While not everyone agrees with this discrimination, there is broader consensus for that type of discrimination than discrimination on the basis of sex or race.

should engineers decide that the most popular content – and not, for instance, the most uplifting, nor the most educational content – should be the most visible content. After all it is an entertainment website, not an educational website. Trustworthiness may not be prioritized to the same extent that, for instance, a news website might. Nonetheless, even an entertainment website cannot entirely bracket the value of trustworthiness. If YouTube algorithms were to promote a documentary that contained defamation or incitement, this would clearly be problematic.

This statement about the proper function of algorithm engineering is prescriptive; however, it also closely corresponds to how online platforms are in fact regulating content. For instance, consider how the Google algorithm is supplemented by the judgments of are called Search Quality Raters. These employees read websites and judge them for trustworthiness: their judgments are then fed into the Google algorithm in order to tweak the results. These Raters are given extensive guidelines (Google 2021) on how to check and correct the results of the Google algorithm. In particular, Google looks for their input with regard to searches on topics where the cost of error is high, such as "Your Money or Your Life" webpages, i.e., webpages about personal finance and healthcare (Google 2021, 10). Here Google emphasizes the importance of getting the rank of these webpages right.

What are the guidelines? What Search Quality Raters must look for, in particular, is E-A-T: Expertise, Authority, and Trustworthiness (Google 2021, 19–20). How are these properties discerned? By traditional markers of prestige and trustworthiness: professional qualifications (J.D., M.D., etc.), awards by accredited bodies (Pulitzer prizes, etc.), signs that a webpage has been edited and reviewed, websites with rigorous editorial and review policies (Google 2021, 20). Search Quality Raters must even be able to judge information pages on whether they have been written by "people or organizations with appropriate scientific expertise" and whether the pages "represent well-established scientific consensus on issues where such consensus exists" (Google 2021, 20).

This is striking: machine learning algorithms were originally intended to learn automatically from the behavior of crowds of users, in a wisdom-of-the-crowds rationale (Masterton and Olsson 2018). However, today, even the most widely used search engine, with privileged access to data of user behavior, must be redirected by specially trained employees in order to make better judgments of trustworthiness and relevance. In other words, the latest iteration of Google search depends on the judgments of two groups of *gatekeepers*: the Raters who decide which results to promote and which results to suppress, and the scientists and professionals on which the Raters base their judgments.

This illustrates how even the latest iteration of the most dominant search engine is explicitly geared towards promoting trustworthiness in its search results (and especially regarding types of query where trustworthiness is crucial). The proper function of algorithm engineering also does not aim at *replacing* human judgment of trustworthiness (regarding competence or honesty), but rather as *implementing* human judgments in contexts characterized by considerable information overload.

7. Democracy, Transparency, and Professionalism

In the *Open Society,* Popper attacks what he calls "utopian social engineering", or how 20th century totalitarian states sought to control behavior of citizens through propaganda, ideology, and other means (Podgórecki, Alexander, and Shields 1996). And the term "trust engineering"[8] inevitably inherits some of the pejorative connotations associated with "social engineering". One could worry whether trust engineering gives engineers too much power, or whether it can be used to legitimize forms of undemocratic social control or manipulation.

Here algorithm engineering simply inherits the difficulties associated with regulating speech. It can be objectively difficult to distinguish between critical discussion and proto-incitement. For instance, say that a scientific criticizes the standard scientific view that intelligence, while it has a genetic component, does not have a genetic component that correlates with "race". How should that criticism be categorized? In principle it such criticism could be viewed as part and parcel of genuine scientific enquiry. However, it could be a willful distortion of evidence, or a foregrounding of certain facts about reality that are of no genuine scientific interest and that can only be of interest for a strategy of "scientific racism" aimed at promoting discrimination towards certain groups of people. Even before the advent of the online social environment it was difficult to regulate such speech acts, and the judgments implemented by algorithms simply inherit such challenges.

[8] I'll use the term "trust engineering" as shorthand for the type of algorithm engineering that aims to safeguard the conditions of trustworthiness in the online social environment.

Such challenges do not constitute an argument *against* trust engineering, merely serve to illustrate how important it is to acknowledge the role that *ethical deliberation* plays in the design of algorithms. Ethical deliberation involves the weighing and choosing a course of action based on competing values (as in, e.g., principilism: Beauchamp and Childress 1979). Manipulation occurs when this deliberation is hidden from public view, and when a small group of individuals decides what a large group of users gets to see.

Can worries about the undemocratic nature of algorithmic speech regulation be circumvented by simply making algorithms transparent and intelligible (e.g. Coglianese and Lehr 2019)? This is a large and lively discussion, but the vision sketched in this chapter gives reasons to be skeptical that mandating transparency and even intelligibility will be sufficient. If algorithms are thought of as shaping an external (online) environment, most individuals are relatively powerless in such a situation and will face strong incentives to simply adapt to that environment. In this way an "accountability gap" emerges between what an engineer is directly responsible for (i.e., the design of an algorithm) and the downstream effects on the landscape of the online environment (Kiester and Turp 2022; Mittelstadt et al. 2016). Algorithm engineers may be entirely transparent about their design choices, and yet the effect could still be a stifling of free and open critical discourse when users remain too beholden to the judgments of trustworthiness generated by the algorithm.

As a concrete example developed more fully in other work (Desmond forthcominga), search algorithms of science repositories can strengthen existing *status biases,* where disproportionate attention is given to individuals based on their perceived social status rather than based on the quality of their work. Such biases are not "controlled" by anybody, but rather are a consequence of evolved cognitive biases. Yet they erect new, invisible barriers erected to some of the values that open science strives for (such as inclusiveness or democracy), and being transparent about the design of search algorithms does not necessarily help. For instance, information regarding the design of PubMed's search algorithm is in the public domain (PubMed is publicly funded), and since the late 2000s, PubMed has relied on machine learning algorithms, where revealed user preferences (what they click on given certain search terms) influences the future ranking of results. Such search algorithms perpetuate biases, and even if users would know about the biased design of such algorithms, that would not prevent them from disproportionately clicking on the first search results (and thus reinforcing the existing biases).

The social engineering involved in algorithm engineering thus can be very closely compared to environmental engineering, where the engineer will intervene on the environment and will modulate the relative fitnesses of different populations. The environmental engineer is not like a breeder: the engineer does not decide for each individual organism whether they get to reproduce or not. Rather, the environmental engineer intervenes on certain structural features in the environment, and then lets natural selection do the work in "judging" organisms in the natural environment. In the same way the algorithm engineer intervenes on certain parameters in the design of the algorithm, and then lets the algorithm do the work in "judging" communicative acts in the online environment.

The preceding considerations point to a more accurate causal structure of how online communications are evaluated, one where both the design of the algorithm as well as the agency of the engineer (i.e., their competence and values) play a role. Expanding on Figure 3, Figure 5 systematizes this structure: if individual A produces the communicative act X, A is not the only the difference-maker for the message to be endorsed or not by B (variable Y). Rather, it is the intervention by the algorithm (search engine, recommendation algorithm) that makes the difference that B is exposed to communicative act X instead of acts X' and X". Thus the crucial difference-maker in this picture is the algorithm (Z), not the person B producing the message. The algorithm is in turn designed by the engineer γ. In this way, the dynamics of trust between A and B in the online environment depends on the trustworthiness of the algorithm engineer.

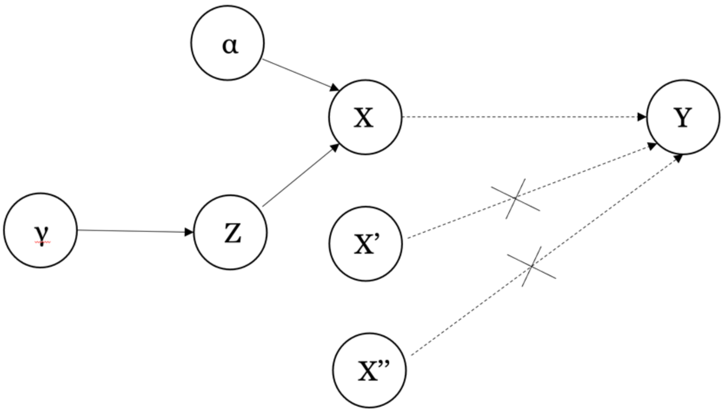

Figure 5: When B must evaluate A's statement X for trustworthiness, not only is A's agency (α) relevant here, but also the design of the algorithm (Z) which has determined that X is presented to B (and not X' or X''), and the agency of the engineer (γ) who has caused Z.

Speaking of the trustworthiness of engineers suggests a different model of regulation of : that of professionalism. Between engineers and users there is at least as much expertise asymmetry as between a physician and a patient, or between a lawyer and a client. Hence, if the model of professional ethics were to be applied to algorithm engineers, this expertise asymmetry would be paired with not just norms of competence but also service ideals (these remarks about professional ethics are developed more in Desmond 2020). On this view, it may not be necessarily problematic that algorithm design is, on a proximate level, an undemocratic process carried out by engineers as along as engineers are guided by a set of values that *is* endorsable as service to broader society.

However, much work would need to be done here, as it remains unclear what values algorithm engineers should aim at in their work. Promoting trustworthiness and conditions for honest communication should likely be one general value, but other, more concrete values would be needed as well, especially those values pertaining to what type of public discourse we wish to promote. Identifying those values would be a difficult and controversial task.

8. Conclusion

It may seem strange or even sinister to talk about “trust engineering”, but it is through algorithms that we perceive the online environment and thus decide whether or not to trust some online communication. Algorithms help determine what communications are *worthy* of our attention, and thus potentially of our trust. They also shield users from communicative acts that are harmful, either to the users themselves or else to communal norms of trustworthiness. This is not without dangers: once the social function of algorithm engineering is acknowledged, algorithm engineering can be misused for purposes of social control, where not only the conditions of trustworthiness are safeguarded, but the deliberative freedom of users actively undermined. However, it is precisely because algorithm engineering can be abused that is important that its value-laden nature is recognized, and that a robust professional ethics for algorithm engineers is developed.

Acknowledgments

I would like to thank Mark Alfano for comments on an earlier draft of the chapter, and David Collins for guiding the chapter through to publication.

References

Aaby, Bendik Hellem, and Hugh Desmond. 2021. “Niche Construction and Teleology: Organisms as Agents and Contributors in Ecology, Development, and Evolution.” *Biology & Philosophy* 36 (5): 47. https://doi.org/10.1007/s10539-021-09821-2.

Alexander, Jennifer Karns. 2020. “A Brief History of Engineering.” In *The Routledge Handbook of the Philosophy of Engineering*, edited by Diane P. Michelfelder and Neelke Doorn, 1st ed., 25–37. New York: Routledge. https://doi.org/10.4324/9781315276502-4.

Alfano, Mark, Amir Ebrahimi Fard, J. Adam Carter, Peter Clutton, and Colin Klein. 2021. “Technologically Scaffolded Atypical Cognition: The Case of YouTube’s Recommender System.” *Synthese* 199 (1): 835–58. https://doi.org/10.1007/s11229-020-02724-x.

Baier, Annette. 1986. “Trust and Antitrust.” *Ethics* 96 (2): 231–60. https://doi.org/10.1086/292745.

Bartling, Sönke, and Sascha Friesike, eds. 2014. *Opening Science*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8.

Beauchamp, Tom L., and James F. Childress. 1979. *Principles of Biomedical Ethics*. Oxford University Press.

Carpenter, Jordan, William Brady, Molly Crockett, Rene Weber, and Walter Sinnott-Armstrong. 2021. “Political Polarization and Moral Outrage on Social Media.” *Connecticut Law Review* 52 (3): 1107–20.

Chamberlain, Geoffrey. 2006. “British Maternal Mortality in the 19th and Early 20th Centuries.” *Journal of the Royal Society of Medicine* 99 (11): 559–63. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1633559/.

Coglianese, Cary, and David Lehr. 2019. “Transparency and Algorithmic Governance.” *Administrative Law Review* 71: 1. https://heinonline.org/HOL/Page?handle=hein.journals/admin71&id=13&div=&collection=.

Covington, Paul, Jay Adams, and Emre Sargin. 2016. “Deep Neural Networks for YouTube Recommendations.” In *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–98. Boston Massachusetts USA: ACM. https://doi.org/10.1145/2959100.2959190.

Desai, Deven R, and Joshua A. Kroll. 2017. “Trust but Verify: A Guide to Algorithms and the Law.” *Harvard Journal of Law & Technology* 31: 2–64.

Desmond, Hugh. 2020. “Professionalism in Science: Competence, Autonomy, and Service.” *Science and Engineering Ethics* 26 (3): 1287–1313. https://doi.org/10.1007/s11948-019-00143-x.

———. 2021a. “Incentivizing Replication Is Insufficient to Safeguard Default Trust.” *Philosophy of Science*, no. 88: 1–12. https://doi.org/10.1086/715657.

———. 2021b. “Adapting to Environmental Heterogeneity: Selection and Radiation.” *Biological Theory*, March. https://doi.org/10.1007/s13752-021-00373-y.

———. forthcoming. “Conserving Gatekeepers in the Open Science Era.” *Synthese.*

———. forthcoming. “Reclaiming Care and Privacy in the Age of Social Media.” In *Virtues and Values in a Changing World*, edited by Anneli Jefferson, Orestis Palermos, Panos Paris, and Jonathan Webber. Cambridge, UK: Cambridge University Press.

Desmond, Hugh, and Kris Dierickx. 2021. “Trust and Professionalism in Science: Medical Codes as a Model for Scientific Negligence?” *BMC Medical Ethics* 22 (1): 45. https://doi.org/10.1186/s12910-021-00610-w.

Desmond, Hugh, and Grant Ramsey, eds. in press. *Human Success: Evolutionary Origins and Ethical Implications*. Oxford, UK: Oxford University Press.

Flink, James J. 1988. *The Automobile Age*. Cambridge, Mass: MIT Press.

Giest, Sarah. 2017. “Big Data for Policymaking: Fad or Fasttrack?” *Policy Sciences* 50 (3): 367–82. https://doi.org/10.1007/s11077-017-9293-1.

Gilman, Michele. 2021. “AI Algorithms Intended to Root out Welfare Fraud Often End up Punishing the Poor Instead.” The Conversation. 2021. http://theconversation.com/ai-algorithms-intended-to-root-out-welfare-fraud-often-end-up-punishing-the-poor-instead-131625.

Google. 2021. “General Guidelines.” 2021. https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf.

Hawley, Katherine. 2014. “Trust, Distrust and Commitment.” *Noûs* 48 (1): 1–20. https://www.jstor.org/stable/43828859.

Kiester, Lucy, and Clara Turp. 2022. “Artificial Intelligence behind the Scenes: PubMed’s Best Match Algorithm.” *Journal of the Medical Library Association* 110 (1): 15–22. https://doi.org/10.5195/jmla.2022.1236.

Koen, Billy V. 2003. *Discussion of the Method: Conducting the Engineer’s Approach to Problem Solving*. Oxford University Press.

Laland, Kevin, Blake Matthews, and Marcus W. Feldman. 2016. “An Introduction to Niche Construction Theory.” *Evolutionary Ecology* 30: 191–202. https://doi.org/10.1007/s10682-016-9821-z.

Masterton, George, and Erik J. Olsson. 2018. “From Impact to Importance: The Current State of the Wisdom-of-Crowds Justification of Link-Based Ranking Algorithms.” *Philosophy & Technology* 31 (4): 593–609. https://doi.org/10.1007/s13347-017-0274-2.

Mitcham, Carl. 2020. “What Is Engineering.” In *The Routledge Handbook of the Philosophy of Engineering*, edited by Diane P. Michelfelder and Neelke Doorn, 11–24. New York: Routledge.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. “The Ethics of Algorithms: Mapping the Debate.” *Big Data & Society* 3 (2): 2053951716679679. https://doi.org/10.1177/2053951716679679.

Mühlichen, Michael, Rembrandt D. Scholz, and Gabriele Doblhammer. 2015. “Social Differences in Infant Mortality in 19th Century Rostock. A Demographic Analysis Based on Church Records.” *Comparative Population Studies* 40 (2). https://doi.org/10.12765/CPoS-2015-03.

Munn, Luke. 2020. “Angry by Design: Toxic Communication and Technical Architectures.” *Humanities and Social Sciences Communications* 7 (1): 1–11. https://doi.org/10.1057/s41599-020-00550-7.

Pinker, Steven. 2018. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. Penguin.

Podgórecki, Adam, Jon Alexander, and Rob Shields. 1996. *Social Engineering*. McGill-Queen’s Press - MQUP.

Ritchie, Hannah, and Max Roser. 2014. “Natural Disasters.” *Our World in Data*, June. https://ourworldindata.org/natural-disasters.

SPJ, (Society of Professional Journalists). 2014. “Society of Professional Journalists Code of Ethics.” https://www.spj.org/ethicscode.asp.

Statista. 2021. “Internet Users in the World 2021.” Statista. 2021. https://www.statista.com/statistics/617136/digital-population-worldwide/.

Twenge, Jean M., Jonathan Haidt, Thomas E. Joiner, and W. Keith Campbell. 2020. “Underestimating Digital Media Harm.” *Nature Human Behaviour* 4 (4): 346–48. https://doi.org/10.1038/s41562-020-0839-4.

Twenge, Jean M., Thomas E. Joiner, Megan L. Rogers, and Gabrielle N. Martin. 2018. “Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time.” *Clinical Psychological Science* 6 (1): 3–17. https://doi.org/10.1177/2167702617723376.

Twitter. 2014. “The 2014 #YearOnTwitter.” 2014. https://blog.twitter.com/en_us/a/2014/the-2014-yearontwitter.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford, UK: Oxford University Press.

Zalasiewicz, Jan, Mark Williams, and Colin Waters. in press. “Anthropocene Patterns in Stratigraphy as a Perspective on Human Success.” In *Human Success: Evolutionary Origins and Ethical Implications*. Oxford, UK: Oxford University Press.