Algorithmic neutrality [1,2]

Milo Phillips-Brown // milopb@ed.ac.uk
University of Edinburgh

Draft as of June, 2024 // comments welcome

Abstract. Algorithms wield increasing control over our lives—over the jobs we get, the loans we're granted, the information we see online. Algorithms can and often do wield their power in a biased way, and much work has been devoted to *algorithmic bias*. In contrast, *algorithmic neutrality* has been largely neglected. I investigate algorithmic neutrality, tackling three questions: What is algorithmic neutrality? Is it possible? And when we have it in mind, what can we learn about algorithmic bias?

In 2005, Adam and Shivaun Raff started a small business: Foundem, a comparison-shopping site similar to Google Shopping. Foundem showed promise. At one point, it was hailed as the UK's best comparison-shopping site. But on June 26, 2006, Google changed its search algorithm, dropping foundem.com from the top three search results to the 70s. By all indications, Foundem's drop in Google's rankings was not due to a drop in quality. Foundem.com still held a top place in Yahoo's and Microsoft's search rankings. But in the search engine optimization industry, it's said that if you want to bury a body, you put it on the second page of Google. And Foundem was no exception. It would not recover from the loss of traffic from Google (Manthorpe, 2018).

In one way, Foundem's story is unremarkable: Foundem alleged that they were victims of *algorithmic bias*, and it's well documented that bias infects search engines and algorithms more generally. For example, in 2017 the EU found that Google's search engine was biased in its own favor; Google Shopping undeservedly enjoyed higher search rankings than rival comparison shopping services, including Foundem (European Commission, 2017). (The result was a €2.42 billion fine.) Search engines are biased in other ways, too. Introna and Nissenbaum (2000) argued that the technical architecture of search engines excludes the voices of the less powerful and less wealthy. Noble (2019) revealed how search engines perpetuate sexism and racism by returning highly sexualized results for queries like 'Black girls'. Beyond search engines, we find bias in predictive policing software that overestimates crime in communities of color (Lum and Isaac, 2016); hiring algorithms that dock qualified female candidates (Barocas and Selbst, 2016); and so on.

In another way, though, Foundem's story *is* remarkable: following Foundem's demotion in Google's search rankings, its founders initiated the 'search neutrality' movement. *Algorithmic neutrality* has received little attention, despite the considerable work that's been devoted to algorithmic bias.

[1] For many conversations and comments on earlier drafts, thank you to Laura Alvarez Jubete, Serena Booth, David Boylan, Carol Brown, Thomas Byrne, medb corcoran, Ray Eitel-Porter, Sina Fazelpour, Nilanjan Das, Kevin Dorst, Dalia Gala, David Grant, Lyndal Grant, Sally Haslanger, Abby Jaques, Matt Mandelkern, Silvia Milano, Carina Prunkl, Bernhard Salow, Jen Semler, Kieran Setiya, Jack Spencer, Charlotte Unruh, Kate Vredenburgh, and audiences at Northeastern University, the Massachusetts Institute of Technology, the Philosophy in an Inclusive Key Summer Institute, the Annual Meeting of the PPE Society, the University of Oxford, the University of Washington, and the University of Wisconsin. Thank you to Ginger Schultheis for extensive and transformative feedback. Thank you especially to Marion Boulicault, Quinn White, and the members of the Jain Family Institute's Digital Ethics and Governance team for years of inspiration, guidance, and encouragement.

[2] This project was supported by a gift from Accenture.

(Algorithmic *fairness* has been a hot topic (Castro, 2019), but the relationship between fairness and neutrality is far from clear—in general and in the case of algorithms. For example, neutrality, as I'll characterize, is a descriptive notion, while fairness is a normative one.)

This paper investigates three questions about algorithmic neutrality: What is algorithmic neutrality? Is algorithmic neutrality possible? And when we have algorithmic neutrality in mind, what can we learn about algorithmic bias?

To my knowledge, these questions have never been systematically asked or answered. While it's often said that algorithms aren't (or can't be) neutral, what's usually meant is simply that the algorithm is biased or unfair (or must be so). Neutrality—in its own right—has not received sustained, dedicated attention.[3] For example, to my knowledge no one has ever stated—much less argued for—an explicit, general characterization of algorithmic neutrality. (There are characterizations of two *specific* forms of algorithmic neutrality, as I discuss just below.) In contrast, explicit, general characterizations and taxonomies of algorithmic bias and fairness abound. These characterizations and taxonomies have proven enormously productive in understanding the nature, possibility, and normative significance of algorithmic bias and fairness. We need the same for algorithmic neutrality.

In addressing the three questions, I will start with a case study: search engines. Search engines warrant special attention because they themselves are remarkable in discussions of algorithmic neutrality. Search neutrality is one of only two specific kinds of algorithmic neutrality to receive extensive public, academic, and legal attention.[4,5]

I answer the three questions in turn, as applied to search engines: in §1, I draw on work about neutrality in science to characterize neutrality in search;[6] in §2, I argue that search neutrality is impossible; in §3, I discuss how to make sense of search bias—given that search neutrality is impossible—and introduce kind a new of bias, *external-values bias*. Much of what I say in §1–3 is stated in terms of the *aim* of a given search engine; in §4, I explore the notion of an aim. In §5, I generalize my discussion of search engines to algorithmic systems of all kinds.

1 What is search neutrality?

To characterize search neutrality, I will first characterize a more general kind of neutrality—algorithmic or otherwise—of which I will argue search neutrality is an instance.

Imagine a governmental agency in charge of distributing vaccines among a country's provinces. The law mandates that vaccines be distributed to provinces so as to maximize lives saved. But, motivated by political allegiances, the agency withholds vaccines from provinces that are governed by a certain party. In so doing, it does not distribute vaccines neutrally. Or imagine that you are assigning grades for participation in a course. You don't keep a record of how much students talk, instead going off of memory. You happen to know which students in your class have parents that donate to your university,

[3] Building on work in philosophy of science, some—such as Dotan (2020), Johnson (fc)—investigate whether algorithms are 'value-free', which these theorists understand as freedom *non-epistemic values* (I explain this notion in §1). 'Neutrality' and 'freedom from non-epistemic values' are often used synonymously by philosophers of science; with science, neutrality and freedom from non-epistemic values may well coincide. But this is not in general the case, as I discuss in §1 and footnote 16. We would be uncharitable, then, to read Dotan and Johnson (and others) as theorizing about algorithmic neutrality *per se*.

[4] The other is net neutrality.

[5] See, among many others, (Grimmelmann, 2010; Crane, 2012; Gillespie, 2014).

[6] In so doing, I build on the work of others, noted in footnote 3, who connect algorithms to this work in philosophy of science.

and—despite your best efforts not to—you subconsciously inflate the participation grades of students whose parents donate. In so doing, you do not assign grades neutrally. Finally, imagine a scientist investigating the effects of smoking on cancer. Her research is funded by a tobacco company, and she selectively ignores evidence that's unfavorable to her funders' financial interests. In so doing, she does not carry out her research neutrally.

To articulate the general kind of neutrality common to these cases, I'll begin with a particular kind of neutrality—neutrality in science[7]—which is standardly characterized along these lines:

> **Neutrality in science**[8]
> Science is neutral only if non-epistemic values play no role in how scientists conduct core scientific practices.

To shed light on algorithmic neutrality, we need only a basic understanding of what neutrality in science amounts to. What distinguishes epistemic from non-epistemic values? Pursuing epistemic values—like empirical adequacy or internal consistency—leads us to truth (Steele, 2012). (Just what empirical adequacy and internal consistency are doesn't matter for this paper.) Pursuing non-epistemic values—like financial interests and political ideologies—doesn't lead us to the truth. What is a core scientific practice? The idea is that neutrality is compromised if non-epistemic values play a role in how scientists engage in *certain* scientific practices (like gathering evidence), but not in others (like applying for grants). The former are core scientific practices; the latter aren't.

So, for example, science isn't neutral in the case of the cancer researcher because a non-epistemic value (her funder's financial interests) plays a role in how she conducts a core scientific practice (gathering evidence).

There are many things one could ask about scientific neutrality: where to draw the line between epistemic and non-epistemic values; what counts as a core scientific practice; or whether *ideal* science is neutral: theorists debate whether science can be neutral, and if so, whether it should be.

These questions aren't the topic of this paper. Rather, what's of interest here is the notion of neutrality at stake in science: why should scientific neutrality require that non-epistemic values play no role in core scientific practices? I suggest that it's because science, as traditionally understood, *aims at truth* (Reiss and Sprenger, 2020). Truth is science's north star. Scientific neutrality therefore demands that core scientific practices are guided only by that star. If non-epistemic values—values that don't lead us to the truth—play a role in core scientific practices, those practices deviate from science's aim. Science is thereby not neutral.

The point, put another way is that this characterization of scientific neutrality rests on the idea that epistemic values are *internal* to the aim of science, while non-epistemic values are *external*. This suggests a general characterization of neutrality—in terms of the aim of a system or practice and values that are either internal or external to that aim. It's this kind of neutrality, I maintain, that's at issue in the examples I gave above. In the vaccine case, the law aims to distribute vaccines to maximize saving lives. When values external to this aim—like political values—play a role in how the agency distributes vaccines, the agency's distribution practice is thereby not neutral. Likewise, your grading policy aims to distribute grades on the basis of participation. When values external to this aim—like financial values—play a role in how you grade, your grading practice is thereby not neutral.

[7] Or, as it's often put, 'value-freedom' in science. See footnote 3 on how value-freedom relates to neutrality.

[8] See for example (Douglas, 2009).

We can, then, characterize search neutrality in terms of the aim of search engines.[9]

What is that aim? The standard answer is *relevance* (Munton, 2022). As Google CEO Sundar Pichai claimed when accused of bias by members of the US congress: 'We provide users with... the most relevant information. And that's our true north' (C-SPAN, 2018). Why think that search engines aim to give relevant results? Because the Internet is vast. When we want to find information or explore a topic online, we often cannot do it on our own, or at least not easily. We often don't know which websites have the information that we're looking for, or which pages on a website would be of help. And so we turn to search engines. We give them queries and, if all goes well, they give us relevant results (I explore the notion of relevance more in §2).

Some search engines aim to give results that aren't *simply* relevant, but that also meet some further criteria—such as having true content. I discuss such search engines in §3.2 and §4, and consider, in §4, what it is for a search engine to have an aim. In the meantime, I'll focus on search engines that aim simply to give relevant results. These search engines deserve special attention because relevance is central to how search engines and search neutrality are normally understood (as this section illustrates). Everything I say about search engines that aim at relevance generalizes to search engines with other aims. (In what follows, I'll often shorten 'search engines that aim to give relevant results' to 'relevance engines', or simply 'search engines' when it's clear that I'm talking about relevance engines.)

We can characterize neutrality for relevance engines like this:

> **Neutrality in search**
> A search engine that aims at relevance is neutral only if values other than relevance play no role in how the search engine ranks pages.

Relevance is internal to search engines that aim at relevance; all other values are external.

For example, imagine a search engine operator that intentionally ranks pages of its own products above those of its competitors' even when its competitors' pages are more relevant. (This doesn't take much imagining; it's exactly what the EU fined Google for doing.) A value other than relevance—the external value of the operator's financial interests—plays a role in how the search engine ranks pages. The search engine is thereby not neutral. Or imagine that an exposé has been published about a scandal involving a politician, Ms. F. The operators of a search engine are partial to Ms. F's party; they bury the exposé low in the search results for the query 'Ms. F scandal'. (Members of the US Congress, on both sides of the aisle, have accused Google of political favoritism (C-SPAN, 2018).) A value other than relevance—the external value of a political ideology—plays a role in how the search engine ranks pages. The search engine is thereby not neutral.

Compare how I've characterized search neutrality with how it's normally characterized by search engine operators, their critics, and scholars alike:

> We [i.e. Google] do get concerns [about political bias] across both sides of the aisle. I can assure you we do this [i.e. deliver search results] in a neutral way. And we do this based on a specific keyword, what we are able to assess as the most relevant information. (Google CEO Sundar Pichai in (C-SPAN, 2018))

> Search Neutrality can be defined as the principle that search engines… should have no editorial policies other than that their results be… based solely on relevance. (Search Neutrality, 2009)

[9]For concreteness, I'll discuss search engines of the traditional style that returns a list of pages. What I'll say also applies to a newer brand of search engine, powered by generative artificial intelligence, that instead directly answers the query.

> Search neutrality… at its heart is some idea that Internet search engines ought… [to] employ 'neutral' search algorithms that determine search result rankings based on some 'objective' metric of relevance. (Crane, 2012, p. 1199)

The similarity between these characterizations and mine is no accident. Implicit in them is what I have made explicit: when we take relevance as the aim of search engines, we can—in keeping with the general, aim-based understanding of neutrality—characterize search neutrality in terms of relevance.

2 Is search neutrality possible?

Search neutrality is not possible. I argue this in §2.1. In §2.2 and §2.3, I develop and rebut objections.

2.1 Search neutrality is impossible

Let's begin with the notion of a *multidimensional* concept. Take intelligence, for example. Jack is quicker-witted than Nashid but worse at math (Kamp, 1975). Is Jack more intelligent than Nashid? There may be no good answer to this question. Certainly, along one *dimension* of intelligence, quick-wittedness, he is more intelligent. Along another dimension, mathematical ability, he isn't. But is he more intelligent, full stop? To answer this question, we need some way to compare Jack's quick wit to Nashid's mathematical acumen. We need to weight the dimensions of intelligence against one another.

So, how should they be weighted? Many theorists—such as Kamp (1975), Sen (1997), and Parfit (2016)—argue that there is no *privileged* weighting for multidimensional concepts. These theorists conclude that multidimensional concepts therefore generate *incomparability*. Jack is neither more nor less intelligent than Nashid, nor are they equally intelligent. They are incomparable with respect to intelligence.

The thesis that multidimensionality leads to incomparability is intuitive and widely accepted, although some contest it.[10] A strong version of the thesis says that whenever there's multidimensionality, there is thereby incomparability; a weaker version says that this is only sometimes so. It is this weaker version of the thesis that I need for my arguments, and I will take it for granted without further argument. (Giving further arguments would go well beyond this paper's scope.) Here in §2.1 I'll assume that multidimensionality leads to incomparability in the cases at hand; in §2.2 I'll consider cases where it might not.

Relevance, I maintain, is a multidimensional concept. To see why, consider a case analogous to Jack and Nashid's. Imagine that you enter 'hurricane' into a search engine.What issues are relevant to the query? There are many candidates: what a hurricane is; whether global warming makes hurricanes more powerful; the human toll of hurricanes, or how that toll is unequal across racial and socioeconomic lines; natural disasters similar to hurricanes, like tsunamis or tornadoes; hurricanes that are often discussed (for example, in the United States, Hurricane Katrina of 2005); how governments in different parts of the world respond differently to hurricanes; stories about particular people or communities who have been affected by hurricanes; etc.

Now imagine two pages, P_1 and P_2. P_1 discusses in detail how hurricanes form, the nature of forest fires, and the history of Hurricane Hortense of 1984 (a hurricane that had a low death toll and no other particularly remarkable features). P_2 only briefly notes how hurricanes form, the nature of

[10] See for example (Dorr et al., fc).

tornadoes, and the history of Hurricane Katrina. I maintain that P_1 is more relevant than P_2 along some dimensions of relevance but not along others.

What are these dimensions? One is a certain kind of *aboutness*. Facts about how hurricanes form, for example, are about hurricanes in a way that facts about, say, the nutritional value of mushrooms are not. Another dimension is a certain kind of popularity. Hurricane Katrina is more often discussed—it's a more popular topic—than Hurricane Hortense, and so pages that discuss Katrina rather than Hortense are, along the dimension of popularity, more relevant. Another dimension still is similarity; pages that discuss tornadoes are more relevant along this dimension than pages that discuss forest fires. The final dimension I'll discuss—there are no doubt more—is a certain kind of informativeness (Roberts, 2012). In particular, the amount of information concerning topics that are about the search term, or that are popular, or that are similar. A given webpage might carry more information concerning hurricanes than another, and in so doing be more relevant than the other along this dimension.

Is P_1 more, less, or equally as relevant as P_2, not merely along a dimension, but full stop? To answer, we need to weight the dimensions of relevance against one another. But, as with intelligence, there is no privileged weighting. P_1 and P_2 are incomparable with respect to relevance.

We can now give a first-pass statement of my argument. A search engine that aims to rank pages simply on the basis of relevance cannot rank P_1 above P_2 (or vice versa). The best it can do is to rank one above the other on the basis of a given weighting of the dimensions of relevance. But which weighting? One can't advert to the aim of relevance itself to answer this question since there is no privileged weighting. In other words, the aim of relevance *underdetermines* which weighting to use in ranking pages, and so underdetermines how to rank pages. Values other than relevance must play a role in determining what this weighting is. Values other than relevance must then play a role in ranking pages. Search neutrality is impossible.[11] (My argument parallels prominent arguments, like Rudner's (1953), that scientific neutrality is impossible on the grounds that the aim of truth underdetermines how to conduct core scientific practices.)

2.2 Objection: the searcher's purposes

An objector might contest my claim that the aim of relevance underdetermines how to weight the dimensions of relevance. The objector's reasoning would begin with an observation I agree with: using search engines is a certain way of inquiring, and in general, what is relevant to an inquiry varies with its *purpose* (Anderson, 1995). This fact is evocatively illustrated by Garfinkel (1981):

> When [infamous bank robber] Willie Sutton was in prison, a priest who was trying to reform him asked him why he robbed banks. 'Well,' Sutton replied, 'that's where the money is.' Clearly there are different… purposes shaping the question and answer. [The priest and Sutton] take different things... to stand in need of explanation. (p. 21)

In the case of search, what's relevant to a query varies with the purpose of the searcher. Imagine one person with a purpose like Sutton's, and another with a purpose like the priest's. Each enters 'Why did Sutton rob banks?' into a search engine. For the first person, the most relevant pages concern issues like the financial windfall, degree of difficulty, and risk of imprisonment in robbing banks. For the second, the most relevant pages concern issues like Sutton's motives, character, or religious background.

[11] For other arguments that neutral search is somehow impossible or incoherent see for example (Grimmelmann, 2010) and (Gillespie, 2014).

We should not simply talk, then, of whether one page is more relevant than another. We should rather talk of whether one page is more relevant than another given a purpose. This is why the statement of my underdetermination argument in §2.1 was only a first pass. My claim was: the aim of relevance underdetermines how to weight the dimensions of relevance. This claim, properly stated, is: *for some purposes*, the aim of relevance underdetermines how to weight the dimensions of relevance given those purposes. On this, my objector and I agree.

But the objector goes further, claiming that the searcher's purposes always and fully determine these weightings. The objector says that once we factor in the searcher's purposes, the multidimensionality of relevance doesn't lead to incomparability.

Let's grant that some purposes do fully determine these weightings; for such purposes, there isn't incomparability in relevance. However, not all purposes are like this. (Remember that I'm *assuming* the (weaker) thesis that multidimensionality leads to incomparability in some cases; what I'm *arguing* is that such cases can arise with search engines.) Imagine a child who searches 'hurricanes' with the purpose of exploring the topic of hurricanes. She is simply curious to learn about hurricanes; she has no particular aspects or facts about hurricanes in mind. Her *exploratory purpose*—and exploratory purposes more generally—doesn't determine how to weight the dimensions of relevance. (It seems to me that the purposes of many searches are partly exploratory, but my argument doesn't rely on this being so.) Search neutrality remains impossible.

2.3 Objection: randomness

An objector could go along with my underdetermination claim, but resist my conclusion that search neutrality is therefore impossible. Specifically, this objector says that if one picks *randomly* among the possible weightings of the dimensions of relevance, then neutrality may well be possible. After all, it's natural to think in general, one is neutral if one picks randomly, and not neutral if one does not.

The objection fails. To see why, consider an analogy. We'll imagine that there's an upcoming presidential election with two Democratic candidates and four Republicans. A certain newspaper can interview only one candidate and it commits to choosing randomly in determining who to interview. But randomly among what? It could pick randomly among individual candidates, giving each candidate an equal—one-in-six—chance to be interviewed. Or it could pick randomly along party lines, giving each party an equal—one-in-two—chance to have a candidate of theirs interviewed. Or it could pick randomly along any other number of lines—gender lines, racial lines, etc.

Let us adopt the objector's view about the connection between randomness and neutrality. If the newspaper picks randomly among individual candidates, it's neutral among individual candidates—but *not* neutral along party lines. (There is a two-in-six chance that a Democrat is interviewed, but a four-in-six chance that a Republican is interviewed.) If the newspaper picks randomly along party lines, it's neutral along party lines—but not among individual candidates. (Each Democratic has a one-in-four chance of being interviewed, but each Republican has a one-in-eight chance.[12]) So, in selecting which candidate to interview, the newspaper can be neutral among individual candidates or it can be neutral along party lines, but it can't be neutral in both ways at once.

As for the newspaper, so too for search engines. Imagine a searcher enters the query 'interview of presidential candidates'; each of the six candidates has been interviewed by a different newspaper; and each newspaper's website has a page dedicated to its own interview. We may imagine that the pages are

[12] If we assume, for simplicity, that the Democrats have the same chance as each other and the Republicans have the same chance as one another.

incomparable with respect to relevance. Can randomness allow the search engine to rank the pages neutrally? No. The search engine can be neutral among individual candidates or it can be neutral along party lines, but it can't be neutral in both ways at once. Search neutrality remains impossible.

3 What can we learn about search bias?

Now that we have in view a characterization of search neutrality, and an argument that it's impossible, there are two lessons we can learn about search bias. I consider them in turn in §3.1 and §3.2.

3.1 Do complaints of search bias rest on a mistake?

The impossibility of search neutrality seems to threaten the significance of search bias. After all, if no search engine is neutral, then every search engine is biased. And as Antony (1993) provocatively asks when discussing bias in epistemology: 'If bias is ubiquitous and ineliminable… what are we complaining about?' (p. 136). Do complaints of search bias rest on a mistake? The answer is *no*—for two reasons.

First, because certain forms of search bias are *not* inevitable. Noble (2019), for example, exposed how search engines delivered highly sexualized results for the query 'Black girls'—even to searchers who were not looking for sexualized material. This is a form of bias, and it is not inevitable. Search engines could simply not return such results. (Indeed, as of today's writing in mid-2024, major search engines no longer return sexualized results for 'Black girls'; they do still return sexualized results for similar queries, like 'Argentinian girls' and 'Thai girls'.)

The same goes for the kind of search bias that Introna and Nissenbaum (2000) identified. Search engines of their day relied heavily on *back-links* to gauge relevance. (A page has a backlink when *another* page links to it.) The idea was that the more back-links a page has, the more relevant it tends to be. While this approach improved on previous approaches to gauging relevance, it had its own problems. Often, pages with more backlinks are less relevant than pages with fewer backlinks. And pages featuring more dominant voices tend to have more backlinks than pages featuring minority voices. So, a search engine that relies too heavily on back-links tends to overestimate the relevance of dominant voices and underestimate the relevance of minority voices. This is a form of bias, and it is not inevitable; search engines need not rely so heavily on back-links.

To illustrate the second way that complaints of bias can be apt, I need to introduce a new notion of algorithmic bias: *external-values bias*. As I've said, A search engine that aims at relevance is neutral only if values other than relevance—values external to the aim of relevance—play no role in how the search engine ranks pages. Correspondingly:

> **External-values bias**
> A search engine that aims at relevance is external-values biased if values other than relevance play a role in how the search engine ranks pages.

Because values other than relevance necessarily play a role in how search engine ranks pages, every search engine is inevitably external-values biased.

What does this mean from the normative point of view? The inevitability of external-values bias is not itself cause for concern. External-values bias is a descriptive notion. It's merely a matter of whether certain values play a certain role, not of whether it's a problem when they do play that role. Complaints of external-values bias can nonetheless have normative force. To see why, we must distinguish two roles that external values may play: they can *complement* relevance or they can *override* it.

Imagine that a search engine ranks a less relevant page above a more relevant one (for some search given some purpose). If it does so because an external value is at play, then that other value *overrides* relevance. For example, consider the case from §1 of the politician Ms. F, who is embroiled in a scandal about which an exposé has been written. To protect Ms. F, a search engine operator that favors her party ranks the page on which the exposé was published far down in search results for the query 'Ms. F scandal'. An external value—a political ideology—overrides relevance. This sort of overriding is exactly what the EU fined Google €2.42 for doing: the EU alleged that Google ranked pages of its own products above those of its competitors even when the former were less relevant than the later; Google's financial interests overrode relevance.

It's *not* inevitable that external values override relevance. In the case of Ms. F, for example, the search engine could simply assign a high rank to the highly relevant page that hosts the exposé. It's no mystery, then, how complaints about values overriding relevance can have normative force.

Because the aim of relevance underdetermines how to weight the dimensions of relevance, external values must play a role in determining how to weight these dimensions. When they play this role, they *complement* relevance. Given the inevitability of external values complementing relevance, just what is the complaint when we complain about external values complementing relevance? The complaint cannot be *that* external values play this complementary role, but rather *which* external values play the role. Such a complaint can have normative bite, since it's not inevitable that one external value rather than another plays this role. And so the question to be answered in adjudicating complaints of inevitable external-values bias is simply: which values should complement relevance?

I will not give a comprehensive answer to this question. Doing so is a major project in its own right, which others, such as Mager (2023), take up. Indeed, there are entire literatures that concern analogous questions.[13] One is in philosophy of science. As I noted in §2, some argue that epistemic values underdetermine how to conduct core scientific practices (and so that scientific neutrality is impossible). If such arguments are right, then it's inevitable that non-epistemic values complement epistemic values in scientific practice. And there is a great deal of work on which non-epistemic values these should be.[14] The other literature spans the various fields—such as value-sensitive design, responsible research and innovation, and design justice—that concern the role that values should play in building technologies in general.

What I will do is make two claims about how to go about answering the question of which values should complement relevance. First, we shouldn't reinvent the wheel; the literatures I just mentioned will be immensely instructive. Second, we should attend to two further questions. One is whether it's legitimate for a given external value—in and of itself—to complement relevance. The other is how pages will be ranked if a given external value does complement relevance.[15]

For example, a search engine operator might elect to weight relevance's dimensions in a certain way because it requires less computational power to implement, and less computational power means lower costs. The complementary external value here is the financial interests of the company. Should this value play a complementary role? To answer, we must ask whether it is legitimate for a search engine operator to weight the dimensions of relevance one way rather than another because doing so

[13] See (Fazelpour and Danks, 2021) for a similar point.

[14] See for example (Longino, 1990), (Anderson, 1995), and (Boulicault and Schroeder, 2021).

[15] This second issue is akin to what's known as 'bias in target variable definition' (Barocas and Selbst, 2016).

would serve their financial interests.

We must also ask how pages would be ranked if the search engine operator's financial interests play the complementary role. To see why, let us add more detail to our case. As we saw in §2.1, popularity is one of the dimensions of relevance. Imagine that Black-run businesses tend to be less popular than white-run businesses. Assigning a high weight to popularity will then result in Black-run businesses occupying lower places in search results than they would if popularity were assigned a lower weight. Imagine further that more computational power is needed if popularity is assigned a high weight. In this case, if the value of the search engine operator's financial interests plays the complementary role, Black-run businesses will tend to be ranked lower than they otherwise would be.

3.2 On the normative significance of a search engine's aim

As I noted in §1, some search engines aim to give results that are not only relevant but that also meet some further criteria. A search engine might aim to give results that also have true content; or are useful; or that satisfy the searcher's preferences; or some combination of these; or some combination of relevance with other notions still. The fact that search engines differ in their aims brings into view something else we can learn about bias: the normative significance of certain forms of bias for a given search engine is beholden to the normative significance of the search engine's aim.

Consider a search engine—a *relevance–truth engine*—that aims to give results that are both relevant and have true content. Moving from relevance engines to relevance–truth engines is a key step in the fight against misinformation. This is because relevance engines spread misinformation. Take the question 'did the Holocaust happen?' A theory that the Holocaust did not happen is relevant to this question, just as the actual history of the Holocaust is. (In general, both p and not-p are relevant to the question of whether p (Roberts, 2012).) So, a relevance engine will rank highly pages that discuss Holocaust denial theories. This is in fact exactly what Google's search engine used to do. In December 2016, for example, the top search result for the query 'did the Holocaust happen?' was a page from the neo-Nazi group Stormfront titled, 'Top 10 reasons why the holocaust didn't happen' (Roberts, 2016b). The relevance of this page was exactly what Google appealed to in explaining its search results:

> 'The fact that hate sites appear in Search results does not mean that Google endorses these views,' said the spokesperson in a statement. According to the company, a site's ranking in search results is determined by computer algorithms using hundreds of factors to calculate a page's relevance to a given query. (Roberts, 2016b)

Because Stormfront's page was relevant to the query, its high search ranking was the correct result, if a disturbing one, for a relevance engine.

In contrast, a relevance–truth engine would not rank Stormfront's site so high, since it's shot-through with falsehoods. Indeed, Google's search engine no longer ranks Stormfront's site so high (or ranks it at all) for this very reason (Roberts, 2016a).

Relevance–truth engines differ from relevance engines in their criteria for neutrality and bias. For example, a relevance engine is, as we know, external-values biased if values other than relevance play a role in how the search engine ranks pages. If the value of truth—a value external to the aim of relevance—plays role in how a relevance engine ranks pages, *the relevance engine is thereby biased*.[16]

[16] Here, the presence of an *epistemic* value—truth—makes for bias (and non-neutrality). This further illustrates the fact, discussed in footnote 3, that neutrality and freedom from non-epistemic values are not—in general—the same thing.

(Indeed, if the value of truth played such a role in a way that lead more relevant pages (like Stormfront's) being ranked below less relevant ones (or not at all), then truth would override relevance.)

In contrast, consider what external-values bias amounts to for a relevance–truth engine. A search engine that aims at relevance and truth is external-values biased if values other than relevance or truth—values external to the search engine's aim—play a role in how the search engine ranks pages. If the value of truth plays a role in how a relevance–truth engine ranks pages, the relevance–truth engine is not thereby biased.

Because how a search engine ranks pages may amount to bias if the search engine has one aim rather than another, some questions about whether bias is worth avoiding amount to questions about what aim a search engine should have. For example, would it be cause for concern that a relevance engine is external-values biased because the value of truth sometimes overrides relevance? This is a question, ultimately, of the extent to which the aim of pure relevance is worth pursuing.

4 What's in an aim?

The aim of a given search engine is central in my accounts of search neutrality and bias. So far, I have worked with a rough and ready understanding of a search engine's aim. This understanding needs refining. We should not talk simply of *the* aim of a search engine, as I have been doing. We can, for example, distinguish between a search engine's *intended* aim, what the search engine's operator intends the search engine to do, and its a *stated* aim, what the operator says that the search engine is doing.

Imagine a search engine operator that claims that its search engine is designed to return relevant results, but which secretly serves the interests of a political party, the S Party. The system's stated aim is to return relevant results. Its intended aim is to serve the interests of the S Party. We can characterize neutrality and bias for each of these aims, and in turn, ask and answer all of the same questions I've posed in the preceding sections about neutrality and bias. For example, the secretly politicized search engine is external-values biased with respect to its stated aim just if values external to this aim—values other than relevance—play a role in how the search engine ranks pages. In contrast, the search engine is external-values biased with respect to its intended aim just if values external to this aim—values other than the interests of the S Party—play a role in how the search engine ranks pages.

Whether a complaint of bias is apt differs depending on which aim is at issue. Imagine that the secretly politicized search suppresses pages with information that would harm the S Party, even when those pages are relevant. An external value is *overriding* relevance—which is internal to the stated aim of the search engine. Members of the public, governments, or political rivals can rightfully complain that the search engine is biased in this way. Not so for a complaint of bias with respect to the search engine's intended aim. In suppressing pages that would harm the S Party, even when the pages are relevant, it's a value *internal* to its intended aim that's playing a role. So, the fact that the interests of the S Party override relevance does not thereby make the search engine external-values biased with respect to its intended aim.

Complaints of bias can also be made with respect to the intended aim. Imagine that the search engine's developers are less diligent in protecting female members of the S Party. As a result, the search engine tends to rank more relevant pages that are harmful to female members of the Party above less relevant pages that don't harm the female members. Values external to the aim of the S Party's interests—sexist values—are playing a role, if implicitly, in how the search engine ranks pages. This sort of bias can engender complaints. Female members of the S Party, it's natural to think, would have standing to complain that the search engine doesn't afford them the protection that it does for male members of the Party. (And this is so even though the secretly politicized search engine ends up delivering more relevant results than it would have had it not been biased in this way!)

The intended and stated aim aren't the only aims we can attribute to a search engine. A search engine might have a *legally required* aim. Or maybe search engines have an *idealized* or *constitutive* aim Munton (2022)—perhaps it's to give simply relevant results, or relevant and truthful results—and that financial or political values are external to that aim. (Relative to this idealized or constitutive aim, then, we could deem a search engine that favors the S Party biased even if its intended *and* stated aims were to favor the S Party.)

My ambition is not to catalogue what kinds of aims a given search engine might have. Rather, I am concerned to point out that different kinds of neutrality and bias are indexed to different aims, and so in evaluating claims of neutrality and bias, we must be explicit about which aim is at issue.

In what follows, I will for simplicity return to talking of *the* aim of a given search engine—or algorithmic system more generally—with the understanding that what I say applies to the system's intended and stated aims, or any other aim that one might attribute to it.

5 Generalizing

This paper is animated by three questions: What is algorithmic neutrality? Is algorithmic neutrality possible? And when we have algorithmic neutrality in mind, what can we learn about algorithmic bias? In §1–3, I answered instances of these questions about search engines. I'll now generalize my answers to algorithmic systems of all kinds.

I've used the aim of a given search engine as the central notion with which to characterize neutrality and bias. For any given algorithmic system, I propose to likewise use its aim as the central notion with which to characterize neutrality and bias.

Consider an algorithmic system for college admissions. Assume that the system aims to rank candidates on the basis of ability. (Recently, a similar system was used—to much controversy—throughout the UK (Coughlan, 2020).)

What is neutrality for the admissions algorithm? The system is neutral only if values other than ability—values external to the system's aim—play no role in how the algorithm ranks candidates.

Is neutrality possible for the admissions algorithm? No. Ability underdetermines how to rank candidates, since ability like relevance—is (presumably) a multidimensional concept. And so values other than ability must play a role in ranking candidates. Neutrality is impossible.

When we have neutrality for the admissions algorithm in mind, what can we learn about bias? First, that certain forms of bias are not inevitable, despite the inevitable non-neutrality of the algorithm. (For example, the UK algorithm just mentioned was biased in that it systematically underestimated the ability of students from publicly-funded schools (Coughlan, 2020); it wasn't inevitable that the algorithm did so.) Second, that a certain form of bias—external-values bias—*is* inevitable. External-values bias can take two forms. A value other than ability might *override* ability. (For example, financial considerations would override ability if the algorithm favored candidates whose parents donated to the school, even when they had less ability.) Or, a value other than ability might *complement* ability. (For example, social justice values would complement ability if, between candidates with incommensurable ability, the algorithm favored candidates from historically disadvantaged backgrounds. This would be a form of affirmative action.) We'd be misguided to allege *that* external values complement ability; what's of concern is *which* external values do so. And, finally, when external values bias arises from external values overriding the aim of the system (ability), the bias is worth avoiding only to the extent that the aim is worth pursuing. (For example, you would embrace this kind external-values bias if you thought social justice considerations should not merely complement, but in fact override considerations of ability.)

More generally, take any given algorithmic system. *What is neutrality for the system?* The system is neutral only if values external to the system's aim play a role in how the system delivers its results.

Is it possible for the system to be neutral? The answer is *no* if the system's aim underdetermines how to deliver the system's results. We have seen how underdetermination arises when the system's aim is multidimensional. Underdetermination has many other sources too, of which I will canvas some. Because underdetermination is pervasive, neutrality is impossible for many—if not most—algorithmic systems.

Underdetermination may arise if a system's aim involves an *arbitrary threshold*.[17] Consider an algorithm for use in a foster care system. (Similar algorithms are in fact used—by, for example, the Department of Human Services in Allegheny, Pennsylvania (Allegheny County, 2022).) The algorithm, imagine, is used to identify whether it's safe for a child in foster care to return to their original family. The aim of the system is to categorize children as at a low, medium, or high risk of being abused if they were to return. How likely must abuse be for a child to be categorized as at high risk? 10%? 20%? 21%? 50%? In other words, what is the threshold of high risk? Likewise, what are the thresholds for low risk and medium risk? The aim of categorizing children at low, medium, or high risk itself underdetermines what these thresholds are.

Underdetermination may arise if the system has *multiple aims*.[18] Imagine an algorithm for use in pre-trial detention decisions in the US judicial system, along the lines of those that are in fact widely used (Angwin et al., 2016). Such decisions are supposed to be based on two factors: if the defendant is released, whether they will commit a crime (likelihood of recidivism) and whether they will fail to appear for a future court appearance (likelihood of flight). Our algorithm assigns a defendant a single risk score that represents their aptness for pretrial detention: the algorithm aims to assign scores on the basis of *both* the likelihood of recidivism and the likelihood of flight. These two aims underdetermine how to assign scores to candidates. Imagine that one defendant has slightly higher risk of recidivism than another while having a slightly lower risk of flight. Which defendant should receive a higher risk score, or should they receive the same score? This is a matter of how to weight likelihood of recidivism against likelihood of risk. To resolve it, we cannot appeal to the dual aims of predicting recidivism risk and predicting flight risk, since these two aims 'disagree' with one another. The aims of predicting recidivism risk and flight risk therefore underdetermine how to assign the single risk score to defendants.

Underdetermination may arise from other sources still. Dotan (2020) and Johnson (fc), among others, show how various scientific practices have direct analogues in algorithmic systems. After all, algorithmic systems often aim to get at the truth—for example, they aim to predict whether someone will commit a crime or fail to appear for a court date. In §2.1, I noted that some argue that the aim of truth underdetermines how to conduct certain scientific practices. Suppose such arguments are sound. Then, Dotan and Johnson show, also sound will be analogous arguments that the aim of truth in certain algorithmic systems underdetermines how that system delivers its results.

Finally: *What can neutrality teach us about bias in a given algorithmic system?* That a certain

[17] Johnson (fc) makes a similar point.

[18] Fazelpour and Danks (2021) make a similar point.

forms of bias—external values bias—is inevitable, but not all forms of bias are. When bias is inevitable, the right complaint to make is not *that* a system is biased, but rather *how* it is biased, and this is a matter of what values should complement the aims of the algorithm system. And in cases where bias arises because values external to a system's aim overrides that aim, bias is only worth avoiding to the extent that the system's aim is worth pursuing.

References

Allegheny County (2022). The Allegheny family screening tool. https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx.

Anderson, E. (1995). Knowledge, human interests, and objectivity in feminist epistemology. *Philosophical Topics*, 23(2):27–58.

Angwin, J., Larson, J., Matthu, S., and Kirchner, L. (2016). Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. *ProPublica*.

Antony, L. M. (1993). Quine as feminist: The radical import of naturalized epistemology. In Antony, L. M. and Witt, C., editors, *A Mind of One's Own: Feminist Essays on Reason and Objectivity*, pages 110–153. Westview Press.

Barocas, S. and Selbst, A. (2016). Big data's disparate impact. *California Law Review*, 104(671):671—732.

Boulicault, M. and Schroeder, S. A. (2021). Trust in science. In Vallier, K. and Weber, M., editors, *Social Trust*, pages 102–21. Routledge.

C-SPAN (2018). Google CEO Sundar Pichai testifies on data collection. https://www.youtube.com/watch?v=WfbTbPEEJxI.

Castro, C. (2019). What's wrong with machine bias. *Ergo: An Open Access Journal of Philosophy*, 6.

Coughlan, S. (2020). Why did the a-level algorithm say no? https://www.bbc.co.uk/news/education-53787203. *BBC*.

Crane, D. (2012). Search neutrality as an antitrust principle. *George Mason Law Review*, 19(5):1199–1209.

Dorr, C., Nebel, J. M., and Zuehl, J. (fc). The case for comparability. *Noûs*.

Dotan, R. (2020). Theory choice, non-epistemic values, and machine learning. *Synthese*, (11):1–21.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

European Commission (2017). Antitrust: Commission fines Google 2.42 billion euro. https://ec.europa.eu/commission/presscorner/detail/en/IP_17_1784.

Fazelpour, S. and Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8).

Garfinkel, A. (1981). *Forms of Explanation: Rethinking the Questions of Social Theory*. Yale University Press.

Gillespie, T. (2014). The relevance of algorithms. In Gillespie, T., Boczkowski, P., and Foot, K., editors, *Media Technologies: Essays on Communication, Materiality, and Society*, pages 167–193. MIT Press.

Grimmelmann, J. (2010). Some skepticism about search neutrality. In Szoka, B. and Marcus, A., editors, *The Next Digital Decade: Essays on the Future of the Internet*, pages 435–459. TechFreedom.

Introna, L. and Nissenbaum, H. (2000). Shaping the web: Why the politics of search engines matter. *The Information Society*, 16(3):98–122.

Johnson, G. (fc). Are algorithms value-free? Feminist theoretical virtues in machine learning. *Journal of Moral Philosophy*.

Kamp, J. W. (1975). Two theories about adjectives. In Keenan, E. L., editor, *Formal semantics of natural language*. Cambridge University Press.

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19.

Mager, A. (2023). European search? How to counter-imagine and counteract hegemonic search with european search projects. *Big Data & Society*, pages 1–13.

Manthorpe, R. (2018). Google's nemesis: meet the British couple who took on a giant, won... and cost it €2.1 billion. https://www.wired.co.uk/article/fine-google-competition-eu-shivaun-adam-raff. *Wired*.

Munton, J. (2022). Answering machines: How to (epistemically) evaluate a search engine. *Inquiry*.

Noble, S. (2019). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Parfit, D. (2016). Can we avoid the repugnant conclusion? *Theoria*, 82(2):110–127.

Reiss, J. and Sprenger, J. (2020). Scientific Objectivity. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

Roberts, J. J. (2016a). Google demotes Holocaust denial and hate sites in update to algorithm. https://fortune.com/2016/12/20/google-algorithm-update/. *Fortune*.

Roberts, J. J. (2016b). A top Google result for the Holocaust is now a white supremacist site. https://fortune.com/2016/12/12/google-holocaust/. *Fortune*.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, 20(1):1–6.

Search Neutrality (2009). Making the case for search neutrality. http://www.searchneutrality.org/search-neutrality.

Sen, A. (1997). *On Economic Inequality*. Clarendon Press.

Steele, K. S. (2012). The scientist qua policy advisor makes value judgments. *Philosophy of Science*, 79(5):893–904.