Bias and the Web

Ricardo Baeza-Yates and Leena Murgai

Abstract Bias is everywhere, sometimes blatantly explicit, but most of the time it's hidden, as it often arises from that which is missing, the gaps in our knowledge or data. In this chapter, we cover what bias is and its different sources: how it arises, persists, feeds back into a system, and can be amplified through algorithms. To exemplify the problem, we use the Web, the largest information repository created by humankind. The first countermeasure against bias is awareness – to understand what is represented—so that we may identify what is not. So, we systematically explore a wide variety of biases which originate at different points on the Web's information production and consumption cycle. Today, many if not all the predictive algorithms we interact with online rely on vast amounts of data harvested from the Web. Biased data will of course lead to biased algorithms, but those biases need not be replicated precisely. Without intervention, typically they are amplified. We start with engagement bias, that is, the difference in rates at which people produce content versus passively consume it. We then move onto data bias: who is producing data on the Web, in what language, and the associated measurement and cultural biases. Algorithmic bias and fairness are intertwined. We discuss the difficulty in defining fairness and provide examples of algorithmic bias in predictive systems. Lastly, we look at biases in user interactions. We discuss how position bias can be mitigated by distributing visuals across results and shared information about other users can lead to different social biases. We discuss how biases continually feed back into the Web and grow through content creation and diffusion.

1 Introduction

Our inherent tendency to favor one thing or opinion over another trickles into every aspect of our lives, creating both visible and latent biases in everything we experience and create. Bias is not new. It has been intrinsically embedded in our culture

R. Baeza-Yates (✉) · L. Murgai
EAI, Northeastern University, Silicon Valley, San Jose, CA, USA
e-mail: rbaeza@acm.org

H. Werthner et al. (eds.), *Introduction to Digital Humanism*,
https://doi.org/10.1007/978-3-031-45304-5_28

and history since the beginning of time. However, thanks to the rise of the Internet and the Web, bias can now impact more people, more swiftly and with less effort than ever before. This has led the impact of bias to become a trending and controversial topic in recent years.

As digital humanism is concerned with development of technology and policies which uphold human rights, democracy, diversity, and inclusion, understanding bias is crucial if we are to build a better world. This understanding is twofold, as it is needed (1) to achieve a fairer society, as we cover next using the Web, and (2) to reflect on the biases within the history of humanism itself. Indeed, humanism is rooted in a White male Christian European conception of the world, which includes ethnic, gender, religious, and geographic biases. Hence, properly addressing these and related biases and their impact, it is an important component in the development of digital humanism, which also addresses these biases, preventing the encoding of neocolonialism in new systems and infrastructure.

Any remedy for bias must first start with awareness, and while awareness alone does not alleviate the problem, it is a necessary first step, regardless of the path forward. Progressive societies accept the existence of social bias. They identify protected features (such as gender, ethnicity, or religion), protected domains (such as healthcare, education, housing, or financial services), and underrepresented groups (such as women or people of color in technology) and use this information to construct solutions, including regulation. They prohibit unfair and systemic biases strategically, via policy and antidiscrimination laws. Some go further in trying to remedy the problem by introducing positive bias through reparations, such as affirmative action programs. All of these should be considered when developing social algorithms that essentially impact people.

For many of us, the Web has become a vital part of how we experience and understand the world. Recent decades have seen unprecedented growth in cloud storage, computer, and infrastructure to take advantage of the accessibility of the Web and manage and make use of the vast amounts of data coursing through it—a trend set to continue. Social progress arguably hinges on the integrity and accessibility of the Web and its contingent systems. As for any tool, with increased use and development, comes increased risk of abuse and misuse. Both can be surfaced by searching for bias.

Bias on the Web reflects our cultural, cognitive, and individual biases and can manifest in subtle ways. This chapter aims to increase awareness of the potential effects of bias in Web usage and content on humanity. People are faced with the ramifications of bias on the Web in the most measurable way when pursuing life goals with outcomes governed in part or largely by algorithms, from loans to personalization (Smith et al., 2016). While the obstacles that result may seem like crucial roadblocks that affect only minorities, representation bias is omnipresent and affects us all, though much of the time we are blissfully unaware of its presence and how it insidiously sways our judgment.

Nowadays, our most prominent communication channel is the Web. Unsurprisingly, then, it is also a place where our individual and collective cognitive biases converge. As social media grows increasingly central to our daily lives, so does the

information about us that can be gleaned from it without our knowledge. For instance, news websites such as the New York Times and Washington Post now use information collected about us from Facebook to decide which news articles we will be most interested in (Pariser, 2011). Search and recommender systems can help filter the vast amounts of data available to us via the Web. They can both expose us to content we may never have otherwise encountered and limit our access to others that we should perhaps be paying attention to. All this makes understanding and recognizing bias on the Web more urgent than ever. Our main aim here is to raise awareness of bias on the Web that can impact us both individually and collectively. We must consider and account for these if we are to design Web systems that lift all of humanity, rather than just the privileged few.

The rest of the chapter is organized as follows. Section 2 introduces the different types of biases, where and how they enter the Web. The following sections cover these biases in more detail: engagement, data, algorithmic, user interaction, and developer biases, respectively. In Sect. 8, we highlight perhaps the most pressing concern: the vicious cycle of bias on the Web. We end the chapter with concluding remarks, further reading, and topics for discussion.[1]

2 Bias

One of the difficulties with bias is that it often results from an absence of information and identifying it requires learning what we do not know. Data gaps are not inconsequential. Data informs how we design the products, services, and systems that support and advance humanity; if you (or people like you) are not in it, the resulting design will not cater for your needs. A swathe of examples can be seen from the gender data gap. Even in the most developed countries, gender bias can be observed in how we design everything: healthcare, housing, offices, and safety features in cars and transportation systems. As a result, women are more likely to be misdiagnosed, seriously injured in car accidents, spend more time traveling, waiting in queues for bathrooms, and be uncomfortably cold at work (Perez, 2019).

The first challenge with bias is how to define and thus measure it. From a statistical point of view, bias is a systematic deviation from the true value (error) caused by an inaccurate parameter estimation or sampling process. But the true distribution or reference value is often unknown. Data is a necessarily biased representation of some truth. Take, for example, classification of people. Someone must make an inherently biased decision about which categories exist and which do not. And the things we measure tend to be proxies for what we really want. In practice, any data relating to an individual is a partial and possibly erroneous representation of who they are.

[1] This chapter is a revised and extended version of Baeza-Yates (2018, pp. 54–61) with additional material from Murgai (2023).

Bias can affect our very perception of the world and people (including ourselves) in opaque and immeasurable ways. One study in 2018 which looked at occupational gender stereotypes in image search found that they were exaggerated when compared with US labor statistics (Kay, Matuszek and Munson, 2015). Participants in the study rated search results higher when they were consistent with occupational gender stereotypes. Simultaneously, they found image search results were capable of shifting people's perceptions of real-world distributions. So, bias on the Web goes both ways. Representational harms, though difficult to measure, are real and play a pivotal role in supporting social hierarchies and hindering social progress (Crawford, 2017).

When all we have is outcomes, how do we measure bias, or rather what do we measure it against? When we look at resource allocation like wealth, it seems natural to make a normative assumption of equality. But more generally, the correct reference value might be less clear and subject to debate. For instance, consider a social variable, such as influence; we would expect there to be some natural variation in the amount of attention individuals garner based on their occupation and this need not be problematic.

Cultural biases can be found in our inclinations based on shared norms and beliefs within our communities. We all belong to some communities and not to others. Our cultural biases mean that we have beliefs or opinions (consciously and unconsciously) about things (including people, from other communities and within our own) in advance of encountering them. As with many other things in life, the remedy for prejudice is education. But the only path to education is via diversity.

Cognitive biases affect the way we think and in turn make decisions. There are many ways in which our thinking and judgment can be impaired. The cognitive bias codex (Weinberg, 2016) provides a helpful categorization of cognitive biases, based on how they manifest. Perhaps the most obvious cause is time pressure; when forced to think fast, we tend to make errors (Kahneman, 2011). The second and third result from unintentionally filtering valuable information, either because there is too much of it, or because it is too complex. Finally, we don't just filter information; we tend to fill the gaps in search of meaning—we imagine what other people might be thinking and lean on stereotypes.

Figure 1 shows how bias is involved in our use and growth of the Web. In the next sections, we explain each of the biases shown in red and classify them by type, beginning with engagement or activity bias resulting from how people use the Web and the implicit bias toward people with limited or no Internet access. The next section addresses bias found in Web data and how it potentially poisons the algorithms that use it, followed by biases created through our interaction with websites and how different types of second-order bias feedback into the Web or Web-based systems. We focus on the significance of these biases and not on the methodological aspects of the research used to uncover them. Further details can be learned by following the references provided herein.

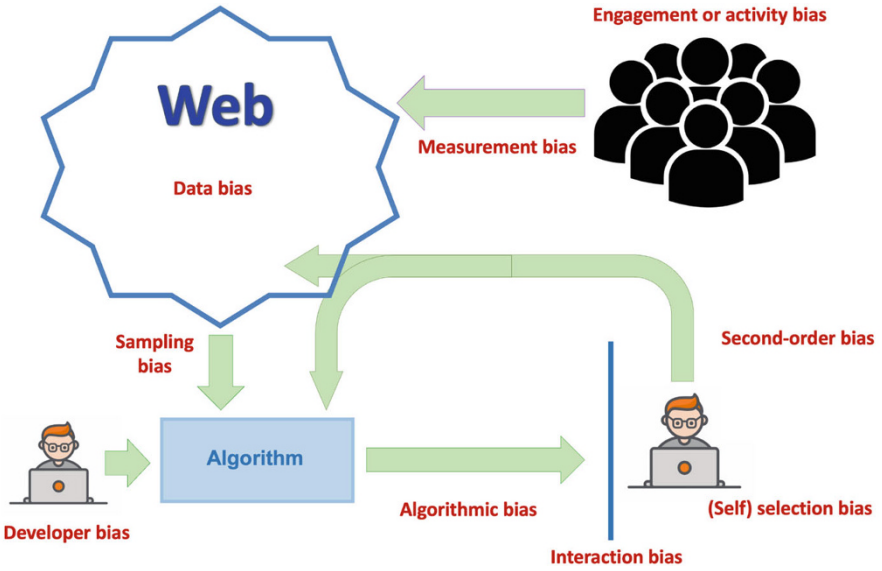


Fig. 1 The vicious cycle of bias on the Web, where the main systemic components are in blue, the biases in red, and their relations in green

3 Engagement Bias: Wisdom of the Few

In 2011, a study of how people followed other people on Twitter found that the top 0.05% of most popular people attracted almost 50% of all the attention (Wu et al., 2011). In other words, half of Twitter users were following a handful of celebrities. Motivated by this fact, we posed the following related question: What percentage of active Web users generated half of the content? That is to say, we did not consider the silent majority that just watches the Web without contributing to it, which is a form of user bias (Gong et al., 2015). We analyzed four datasets, and the results surprised us (Baeza-Yates & Saez-Trumper, 2015, pp. 69–74).

In a small Facebook dataset from 2009, we found that 7% of users generated half the content. In a larger dataset of Amazon reviews from 2013, we found the number to be just 4%. In a very large dataset of Twitter from 2011, the result was even lower, 2%. Finally, the first half of English Wikipedia was written by 0.04% of the registered editors. This indicates that only a small percentage of all users contribute to the Web and the notion that it represents the wisdom of the overall crowd is far from the truth. This is related to Nielsen's 90-9-1 participation rule that states that 1% of the users create content, 9% engage with it (say commenting or doing liking posts), and 90% just lurk (Nielsen, 2006). We also studied the dynamics of these values, finding that at least in Wikipedia, the percentage has increased in the last years as shown in Fig. 2 (courtesy of Diego Saez-Trumper).

A more recent study (Lazovich et al., 2022) looking at engagement on Twitter found similar results that around 90% of people engaged passively. Engagement

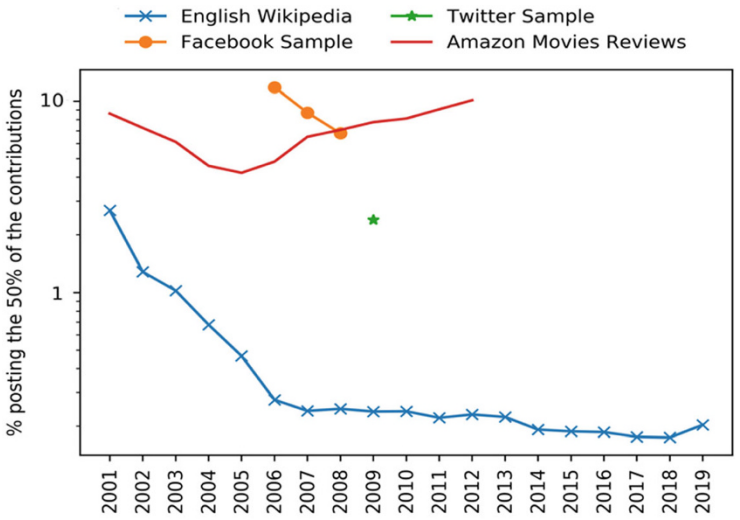


Fig. 2 Dynamics of the percentage of active users that create 50% of the content for four datasets

types that required clicking (likes and author profile clicks) were done by half the population, while retweets, replies, and quote tweets involved the top 70, 80, and 90th percentile of the population, respectively. The top 1% of authors received 80% of the views.

Some remarks on our findings. First, it did not make sense that just 4% of the people were voluntarily writing half of the Amazon reviews. We sensed something else at play. A month after presenting our work, our hunch was confirmed. In October 2015, Amazon started a crusade against paid fake reviews, suing 1000 individuals on a freelance service marketplace accused of writing them (Wattles, 2015). This crusade has continued until today. Our analysis also found that if we considered only the reviews that some people found helpful, the percentage reduced to 2.5% and that there was a correlation between helpfulness and a proxy measure for text quality. Second, although the case of Wikipedia is the most biased, it is a positive example. The 2000 people involved in the start of English Wikipedia probably triggered a snowball effect that helped it become what it is today. Indeed, bias is a requisite in creating anything from nothing.

Zipf's minimal effort law (Zipf, 1949) states that many people do a little while few people do a lot, which may help explain a big part of the engagement bias. However, economic, and social incentives also play a role. For instance, Zipf's law can be seen in most Web measures such as number of pages per website or number of links per Web page. In Fig. 3, we show an example where the Zipf's law is clearly visible on the right side of the graph (the steep line). However, at the beginning, there is a strong social force, the so-called shame effect, which makes the slope less negative. This illustrates that many people prefer to exert the least amount of effort, but most people also need to feel they did enough to avoid feeling ashamed of their

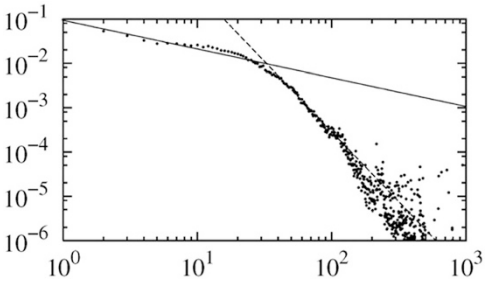


Fig. 3 Shame effect (flatter line) against minimal effort (steeper line) on the number of links in UK web pages, where the *x* axis represents the number of links while the *y* axis is the relative frequency. The intersection is between 12 and 13 links, the average is 18.9, and the exponents of the power laws are 0.7 and 3.6, respectively

work (Baeza-Yates et al., 2007). These two effects are common characteristics of the Web. Notice that data on the far right probably comes from pages written by software, not people.

Finally, as Herbert Simon said, "a wealth of information creates a poverty of attention." Hence, engagement bias generates the *digital desert* of the Web, that is, the Web content no one ever sees (Baeza-Yates & Saez-Trumper, 2015, pp. 69–74). A lower bound comes from the Twitter data where they found that 1.1% of the tweets were posted by accounts without followers! From usage statistics of Wikipedia, we got an upper bound: 31% of the articles added or modified in May 2014 were never visited in June. The actual number likely lies in the first half of the 1–31% range.

In this case, bias can also have advantages. Thanks to engagement bias, all levels of caching are very effective on the Web, making the load on websites and the Internet traffic much lower than it could potentially be.

4 Data Bias

Like people, data quality is heterogeneous and therefore, to some extent, biased. People working in government, university, and other institutions that disseminate information usually publish data of higher quality and attempt to address bias through peer review. Social media data on the other hand is much larger, much more biased, and without doubt, of lesser quality on average. That said, the number of people who contribute to social media (an important subset of Web data) is probably at least one order of magnitude more than those in information-based institutions. Thus, social media produces more data with greater variance in quality, including high-quality data (for any definition of what quality is).

A great deal of bias comes from users' demographics. Internet access and use is, of course, correlated to historical, geographical, economic, and educational factors. These dimensions correlate to other characteristics, having a ripple effect where bias

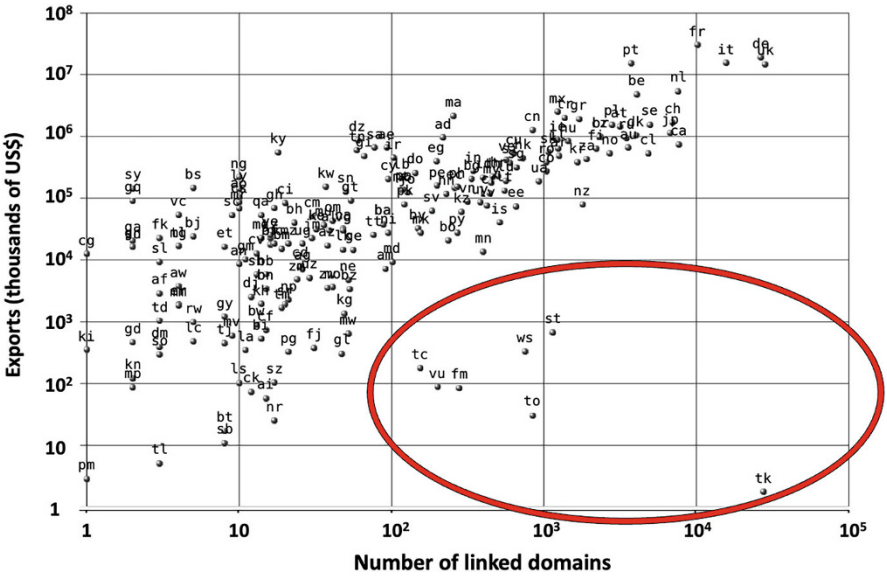


Fig. 4 Economic bias in out links for the Web of Spain (adapted from Baeza-Yates et al., 2006, p. 16)

taints all areas. For instance, it is estimated that over 60% of the top ten million websites (by traffic rankings) are in English (Bhutada, 2021), while the percentage of native English speakers in the world is about 5% (this increases to 13% if we include all English speakers). But is this the correct reference value against which to measure this bias? We could instead use the native language of all people with access to the Internet, where English was almost 26% in 2021 (Statista, 2021). Alternatively, we could consider the percentage of English text on the Web, which might be closer to 30%. At best we still have a bias factor of 2, that is, English websites are twice as prevalent among the best websites as they are among all websites.

Bias can also be found in the link structure of the Web. In Fig. 4, we present a scatter plot showing the total value of Spanish exports to a given country against the number of links from the Web of Spain to the same country (Baeza-Yates et al., 2006, pp. 1–41). Countries in the red circle are outliers; they sold their domain rights for other purposes, such as the Federation of Micronesia, fm, for radio. Discarding those countries, the correlation is over 0.8 for Spain. In fact, the more developed the country is, the higher the correlation, ranging from 0.6 in Brazil to 0.9 in the UK (Baeza-Yates & Castillo, 2006). This does not prove causation, but it is a strong indication of the influence of economy in the link structure of the Web.

What about the representation of women? Consider Fig. 5, which shows the fraction of biographies of women in Wikipedia across history (Graells-Garrido et al., 2015, pp. 231–236). The low fraction of biographies could be explained by the systemic gender bias existing throughout human history (Wagner et al., 2015,

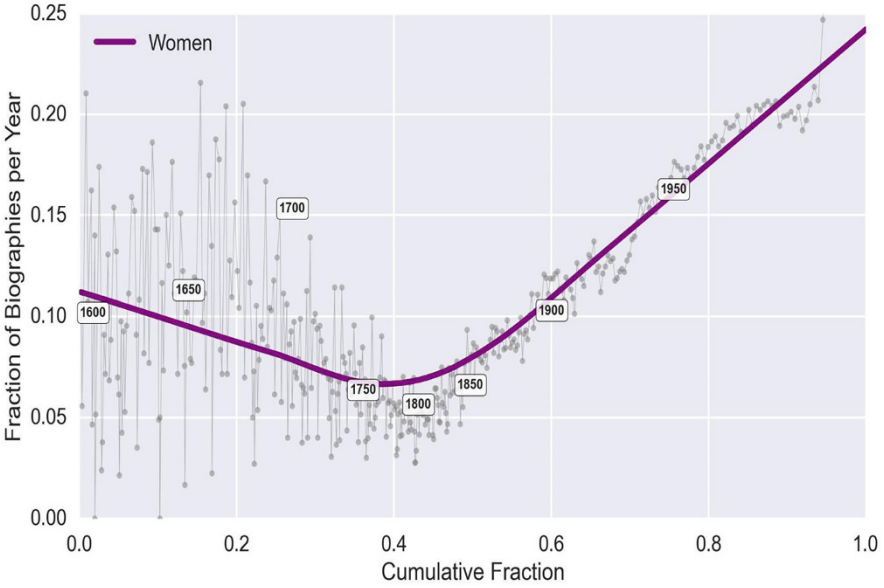


Fig. 5 Accumulated fraction of women biographies in Wikipedia (Graells-Garrido et al., 2015)

pp. 454–463), while the shape seems to change around the French revolution. However, there is an additional underlying factor hiding a deeper bias which is revealed when we look closer at how this content is generated. In the biographies category, less than 12% of the Wikipedia editors are women! In other categories, such as geography, the bias is even worse, falling to a measly 4%. That said, since the percentage of public female editors is just 11%, bias in the biographies category might be viewed as positive rather than negative bias. Keep in mind these values may also contain bias, as not all Wikipedia editors publish their gender; thus, females might be underrepresented in the data as they may prefer not to inform this.

An additional source of data bias is Web spam, a well-known human-generated malicious bias that is difficult to characterize but is motivated by economic incentives. This might be similarly categorized as near-duplication content, like mirrored websites, that represented about 20% of static web content 20 years ago (Fetterly et al., 2003, pp. 37–45).

Since most biases are hard to measure, their effects on predictive algorithms that use machine learning are difficult to understand. First, as Web data represents a biased sample of the population, studies based on social media may have a large error (which we can be sure is not uniformly distributed). Second, the results cannot be extrapolated to the rest of the population for the same reason. As an example, consider the polling errors for past US presidential elections (Mediative, 2014), though online polls performed better than live polls. Third, other sources of error are

biased data samples (e.g., due to selection bias) or samples that are too small for the problem at hand (Baeza-Yates, 2015, pp. 1093–1096).

4.1 Information Bias

Of particular concern in relation to social media platforms are the abundance of bots, disinformation, and fake content that seem to spread faster than real content (Lazer et al., 2018, pp. 1094–1096). In 2020, researchers at CMU analyzed around 200 million tweets related to COVID, over the first 5 months of the year, showing that almost half (45%) of them were by bots (Young, 2020). Of the 50 most influential retweeters, 82% were bots. Among COVID disinformation on social media were conspiracy theories, blaming the outbreak on the introduction of fifth-generation mobile network services. These are believed to have led to 5G towers being burned down in England.

In 2017, Facebook's Ad Manager claimed to reach 41 million 18- to 24-year-olds in the USA, while census data revealed there were only 31 million people of that age group (O'Reiley, 2017). Facebook's role in political and humanitarian crises have been well documented in the media. We've seen targeted misinformation leading up to the Brexit referendum in the UK and US presidential elections in 2016 that led to shocking results (Cadwalladr, 2019) and the more recent insurrection in the US Capitol. Most disturbingly, we have seen concerns around Facebook's engagement optimizing algorithms contributing to social polarization with deadly consequences, especially in regions of ongoing conflict. Amplification of hate speech and incitement of violence on the platform have been implicated in the genocide of Rohingya Muslims in Myanmar and mob violence and crackdowns on independent reporting on Tigray in Ethiopia (Hale & Peralta, 2021) where the deadliest civil war of the twenty-first century rages on (Naranjo, 2023).

4.2 Biases in Language

Perhaps there is no better means to illustrate the intersection of statistical, cultural, and cognitive biases than through language. There are around 7100 living languages in the world, though this number is dwindling with time. They can differ vastly in both their vocabulary and structure. As we've seen, there is no doubt that English is drastically over-represented in language data. What might the disadvantages of over-representing this language, culture, and cognitive universe be?

Research has shown that the languages we speak are related to our cognitive ability in perception tasks (Boroditsky, 2017). For instance, Pormpuraawans (people of an Aboriginal Australian tribe) describe space and time using cardinal directions (north, east, south, west) and consistently order time from east to west. Western languages such as English tend to use more egocentric approaches to describe

position, left, right, in front, behind, and order time from left to right. It should perhaps come as no surprise that Pormpuraawans have superior knowledge of spatial orientation (Boroditsky & Gaby, 2010, pp. 1635–1639). There are more examples; a community in Papua New Guinea who speak Yupno imagine slopes in flat areas (consistent with the valley in which they reside) to describe position. Bardi speakers, from Kimberley in Australia, describe directions as being with or against the tide (Carylsue, 2016). Language has even been known to affect our ability to perceive colors (Winawer et al., 2007, pp. 7780–7785).

In some European languages, such as Spanish and Portuguese, when accidents happen, it is the grammatically correct convention to say, for example, "the glass fell," "el vaso se cayó." In English, it is an accepted convention to say, "Ricardo dropped his glass," regardless of intent. The fact that English speakers take a much more blame-oriented approach in describing mishaps means they are much more likely to remember who was involved (Fausey & Boroditsky, 2008) (or rather accountable, since this is in English). In hindsight, all this makes sense. Language develops in response to the need to describe things. In turn, having words to describe things drives said things into existence, improving our cognitive ability to perceive them.

But language doesn't just affect our cognitive ability; it also shapes our perceptions. In many languages nouns have explicit gendered associations, and some interesting results can be found by comparing languages that ascribe the opposite gender to the same noun. For instance, in Spanish, bridges are masculine, "el puente," while in German, they are feminine, "die Brücke." Researchers have shown that the gender ascribed to a noun can affect the way we imagine them. Indeed, Spanish speakers use more stereotypically masculine words to describe bridges, strong and long, while German speakers use more stereotypically feminine words, beautiful and elegant (Boroditsky et al., 2003, pp. 61–79).

From masculine versus feminine to good versus bad. Given what we have discussed, questions around the connections between gender representation in language and sexism in culture naturally follow (Pitel, 2019). In 2016, the Oxford English Dictionary was publicly criticized for employing the phrase "rabid feminist" as a usage example for the word rabid (O'Toole, 2016). The dictionary included similarly sexist common usages for other words like shrill, nagging, and bossy. A decade before this, historical linguists observed that words referring to women undergo pejoration (when the meaning of a word deteriorates over time) far more often than those referring to men (Trask, 2007; Shariatmadari, 2016). Take, for example, the words mistress (once simply the female equivalent of master, now used to describe a woman in an illicit relationship with a married man), madam (once simply the female equivalent of sir, now also used to describe a woman who runs a brothel), hussy, governess, and the list goes on.

And finally, an example that relates to ordering. In Menominee (Corn Jr, 2019), a Native American language whose roots lie in Wisconsin, people also take a less egocentric approach to describing their interactions and relationships with others, placing themselves after the animate about which they are talking. Both culturally

and in language, they place an emphasis on respect not just for people but all living things, putting others ahead of oneself.

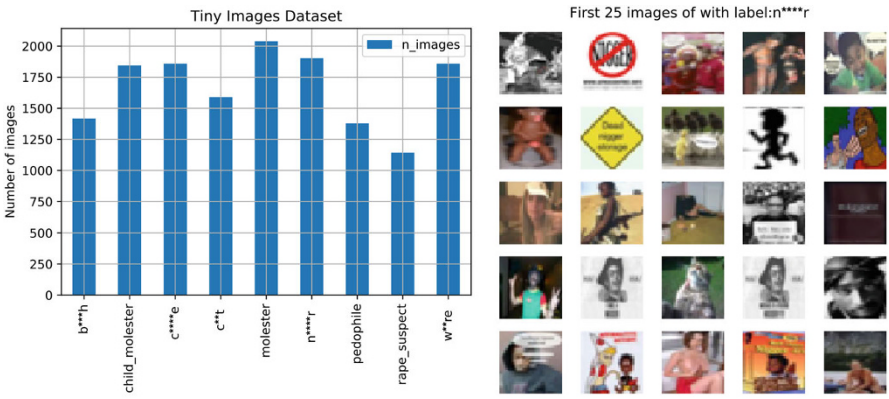
4.3 Bias in Visual Data

Bias in visual records infiltrate the data even before it's been uploaded to the Web, through measurement bias. Capturing likeness in images involves determining the optimal balance of colors to use in each composition. Since its invention, film has been optimized for Caucasian skin. Kodak famously used Shirley cards (Del Barco, 2014) as a standard against which to calibrate colors. It wasn't until the late 1970s, after accusations of racism, that Black, Asian, and Latina Shirleys were added to the reference cards. Today's cameras come with plenty of technology built in to help us take better pictures which we hope is better, but that technology too is imbued with similar biases. Digital cameras assume Asians are blinking (Rose, 2010) and in low light still calibrate to lighter regions to define the image, focusing on White subjects while ignoring darker skin tones (Cima, 2015).

Regarding bias in data quality, we have discussed the good and the bad, now for the ugly. Figure 6 shows some results of an ethical audit of several large computer vision datasets (developed for benchmarking models) in 2020. Researchers found that TinyImages[2] contained racist, misogynistic, and demeaning labels with corresponding images and it was not alone (Prabhu & Birhane, 2021).

The dataset has since been retracted but the problem, unfortunately, does not end there. Datasets used to train and benchmark, not just computer vision but natural language processing tasks, tend to be related. TinyImages was compiled by searching the Web for images associated with words in WordNet (a machine readable, lexical database, organized by meaning, developed at Princeton), which is where TinyImages inherited its labels from. ImageNet (Deng et al., 2009, pp. 248–255) (widely considered to be a turning point in computer vision capabilities) is also based on WordNet, and Cifar-10 and Cifar-100 were derived from TinyImages.

[2] A dataset of 79 million 32 × 32-pixel color photos compiled in 2006, by MIT's Computer Science and Artificial Intelligence Lab, for image recognition tasks.



(a) Class-wise counts of the offensive classes (b) Samples from the class labelled n★★★★r

Fig. 6 Results from the 80 million TinyImages dataset (Prabhu & Birhane, 2021). (**a**) Class-wise counts of the offensive classes, (**b**) samples from the class labeled n****r

5 Algorithmic Bias and Fairness

Today, many if not all the predictive algorithms we interact with online rely on vast amounts of data harvested from the Web. It is scraped from social media, news, product reviews, codebases, and beyond. Algorithmic bias in our case refers to bias that is contributed by the algorithm itself and is not present in the input data. Of course, if the input data is biased (which it is), and the model is calibrated well, the output of the algorithm will reflect that bias; but existing biases in training data can be both amplified and reduced by an algorithm [see, e.g., Kleinberg et al. (2017), Chouldechova (2017)]; the latter is harder to achieve. Making better decisions by identifying, estimating, and accounting for biases requires expertise and ongoing investment, and market forces need not always align with public interests in fairness.

Even if we could detect all possible biases, deciding how an algorithm (or indeed any decision process) should proceed to be fair is in general very difficult. People disagree on controversial issues because the optimal decision is subjective and there are trade-offs. Perhaps the law can guide us here? We've already spoken about protected features and domains; these provide information about the types of problems where we should pay special attention. But how do we decide which trade-offs are acceptable and which are not? Anti-discrimination laws can address both direct discrimination or disparate treatment (making decisions based on legally protected features) and indirect discrimination or disparate impact (policies that disproportionately affect protected groups).

Just as the meaning of fairness is subjective, so is the interpretation of law. At one extreme, anti-classification holds the weaker interpretation that the law is intended to prevent classification of people based on protected characteristics. At the other extreme, anti-subordination principles take a stronger stance, that is, anti-discrimination laws exist to prevent social hierarchies, class, or caste systems, and

legal systems should actively work to eliminate them where they exist. An important ideological difference between the two schools of thought is in the application of positive discrimination policies. Under anti-subordination principles, one might advocate for affirmative action as a means to bridge gaps in access to employment, housing, education, and other such pursuits that are a direct result of historical systemic discrimination against particular groups. A strict interpretation of the anti-classification principle would prohibit such positive bias. Both anti-classification and anti-subordination ideologies have been argued and upheld in landmark cases in the USA.

Perhaps somewhat reassuringly (if only for its consistency), it turns out there are multiple seemingly reasonable definitions of fairness of classifiers which cannot be satisfied simultaneously except in some degenerate cases (Chouldechova, 2017). By Aristotle's definition of fairness (i.e., like cases must be treated alike), deterministic classification is inherently unfair and to resolve this problem in classification, predictions must be randomized (Dwork et al., 2012, pp. 214–226). Interestingly, scholars have shown that privacy concerns are not unrelated to fairness. Note that in both cases we are concerned about protecting certain features. Much like fairness, defining privacy is not a trivial problem; however, it is considered a solved one (Kearns & Roth, 2019). The widely accepted definition of privacy is named differential privacy. It turns out that the solution to the problem of privacy involves adding just the right amount of noise to obfuscate the protected information (Dwork, 2006). The problem of how to define fairness is yet unsolved, though experts predict it will be in the next decade or so (Kearns & Roth, 2019).

In practice, bias from data, and that added by the model, can be hard to separate from a causal perspective. Commercial model developers often expose their model through an API that returns predictions but do not share their training data which has a significant impact on what representations the model has learned. The reality is that choosing training data is a modeling decision. Understanding your distribution of errors through thorough testing and accounting for biases, accordingly, is just responsible modeling. The latter is not currently a requisite for deployment, though the expectation is that regulation of AI will evolve over time and hopefully catch up with other more regulated industries that use predictive modeling at scale such as finance.

So why do models amplify biases in training data? Often at the root of the problem is over-representation of some groups and underrepresentation of others. If one demographic group dominates the data (which is often the case), in the absence of sufficient information for other groups, the algorithm is unable to converge (d'Alessandro et al., 2017, pp. 120–134; Kamishima et al., 2012). Interestingly, this behavioral phenomenon is exhibited not just by models but people too. The term exnomination is well known among those who study culture. It is used to describe the phenomenon of the *default* social class. Members of exnominated groups are privileged because of being the "norm." They have advantages that are not earned, outside of their financial standing or effort, that the "equivalent" person outside the exnominated group would not.

Exnominated groups are catered for by every store, product, service, and system, with preferential access and pricing. They see themselves represented more often overall and more often in a positive light. They are not subject to profiling or stereotypes and more likely to be treated as individuals rather than as a representative of (or as exceptions to) a group. They are more often humanized, more likely to be given the benefit of the doubt, treated with compassion and kindness, and, thus, recover from mistakes. Exnominated groups are less likely to be suspected of crimes; more likely to be trusted financially; have greater access to opportunities, resources, and power; and are able to climb financial, social, and professional ladders faster. The advantages enjoyed by exnominated groups accumulate over time and compound over generations.

In his book *White* (Dyer, 1997), Richard Dyer examines *whiteness* in visual media over five centuries, from the depiction of the crucifixion to modern-day cinema. In many ways, bias on the Web is a living testament to the endurance of the British Empire, through both preservation and continued amplification of its image, language, and culture. Any algorithm trained on Web data, without intervention, will invariably favor White, English-speaking men, to the disadvantage of most of humanity.

So how might we intervene to mitigate bias from Web technology? Well, there are three points at which one should measure and thus could mitigate bias. The first and perhaps most obvious is improving data quality, for example, carefully curating data with diversity in mind. The second attacks the problem with more careful definition of success or objective in training, for example, introducing penalties for undesirable behavior or model constraints based on carefully considered definitions of fairness. Finally, we must monitor model output. One might try to mitigate risk at the end point when a prediction is produced taking countermeasures for cases where we understand our model to be vulnerable.

5.1 Bias in Language Modeling

In 2016, research showed that word embeddings (vector representations of words) generated from news corpora learn biased she-he analogies, e.g., nurse-surgeon or diva-superstar instead of queen-king (Bolukbasi et al., 2016). Why might algorithms exacerbate gender bias? Quick research shows that about 70% of influential journalists are men even though at college age, the gender proportions are reversed. So, algorithms trained on news articles have learned patterns in text developed with demonstrable and systematic gender bias. Other works show that many other cultural and cognitive biases are at play (Saez-Trumper et al., 2013, pp. 1679–1684).

A year later, researchers showed that Google Translate contained similar gender biases (Caliskan et al., 2017, pp. 183–186). They found that "translations to English from many gender-neutral languages such as Finnish, Estonian, Hungarian, Persian, and Turkish led to gender-stereotyped sentences." So, for example, when they translated Turkish sentences with genderless pronouns: "O bir doktor. O bir

hemişre." the resulting English sentences were: "He is a doctor. She is a nurse." They performed these types of tests for 50 occupations and found that the stereotypical gender association of the word almost perfectly predicted the resulting pronoun in the English translation.

Proposals for reducing gender bias include creating more gender balanced data (Costa-jussà et al., 2020, pp. 4081–4088) and mitigating gender bias by transforming embeddings to account for differences in the gender subspace (Bolukbasi et al., 2016). Google opted to intervene at the prediction stage for translations between English and a limited set of just five languages (French, Italian, Portuguese, Spanish, and Turkish), returning both masculine and feminine translations (Kuczmarski, 2018). Google's Natural Language API for sentiment analysis was also found to have problems. In 2017, it was assigning negative sentiment to sentences such as "I'm a Jew" and "I'm a homosexual" and "I'm black"; neutral sentiment to the phrase "white power" and positive sentiment to the sentences "I'm Christian" and "I'm Sikh." In reality, prejudice is, so deeply embedded in language that creating algorithms trained on it that are not is far from trivial.

Bleeding edge developments in language modeling have been focused on conversational capabilities. There is of no doubt that the technology is impressively human sounding, but it also presents some problems for those of us concerned about bias. If machine-written content floods our information ecosystem, what happens to human voices? Chief among model weaknesses is what's described as its ability to hallucinate (a bad metaphor for making a mistake), that is, fabricate expert-sounding, but patently false, prose on complex topics (Hartsfield, 2019). The model is easy to trip up since it cannot reason and does not comprehend. For instance, at the time of writing, ChatGPT was unable to do simple arithmetic, if you ask it to switch the symbols for addition and multiplication first.

There are wider concerns around large language models, specifically their computational inefficiency and corresponding environmental costs (Weidinger et al., 2021). GPT-3, for example, is a model composed of a whopping 175 billion parameters. The costs of building and using this technology are significant when compared to current resources like Google or Wikipedia. Separating fact from fiction is an important milestone if this technology is to be anything more than a rather expensive stochastic parrot (Bender et al., 2021, pp. 610–623) that writes well but needs to be fact checked. Wasting resources does not happen only during training these models but also when billions of people use them as a leisure tool.

5.2 Bias in Computer Vision

In 2015, Google Photos had labeled a photo of a Black couple as gorillas. It's hard to find the right words to describe just how offensive an error this is, but perhaps considering TinyImages, it is not all that surprising. It demonstrated how a machine, carrying out a seemingly benign task of labeling photos, could deliver an attack on a person's dignity.

In 2018, research auditing several popular gender classification packages from IBM, Microsoft, and Face++ showed shocking disparities in performance that depended on both the skin color and gender in sample images (Buolamwini & Gebru, 2018, pp. 1–15).

In 2020, a generative model designed to improve the resolution of images converted a pixelated picture of Barack Obama into a high-resolution image of a Caucasian man (Truong, 2020). If facial recognition technology fails on even the most recognizable faces like Oprah Winfrey, Michelle and Barack Obama, and Serena Williams, what hope do the rest of us have of not being erased by systems that literally can't see us?

5.3 Bias in Recommendations

A major cause for concern is targeted advertising which is now par for the course even in protected domains. In 2013, a study found that Google searches were more likely to return personalized advertisements that were suggestive of arrest records for black names than white, regardless of whether such records existed or not (Sweeney, 2013). This doesn't just result in allocative harms for people applying for jobs; it's denigrating. In 2015, a study showed that women were six times less likely to be shown adverts for high-paying jobs by Google (exceeding $200 K) (Spice, 2015). In 2022, Facebook was fined for using legally protected attributes to target advertisements for housing.

Regarding geographical bias in news recommendations, large cities or centers of political power will naturally generate more news. Hence, if we use standard recommendation algorithms, most people will likely be reading news from the capital and not from the place where they live. Considering diversity and the location of the user, we can give a less centralized view that also shows local news (Graells-Garrido & Lalmas, 2014, pp. 231–236).

An extreme example of algorithmic bias is tag recommendations. Imagine a user interface where you upload a photo, add various tags, and then a tag recommendation algorithm suggests tags that people have used in other photos based on collaborative filtering. You choose the ones that seem correct, and you enlarge your set of tags. This seems like a nice idea, but you won't find this functionality in a website like Flickr. The reason being that the algorithm needs data from people to improve; but as people use recommended tags, they type fewer original ones. They take from the pile without contributing. In essence, the algorithm performs a prolonged harakiri. So, to create a healthy folksonomy (tags made only by people), we should not recommend tags. But we can use these recommended tags to search for similar images by using related (human-produced) tags. Though as we have seen, our ability to find similar images is limited by bias in computer vision technology.

Another critical class of algorithmic bias in recommender systems is related to what items are shown or not shown. This bias affects the user interaction, and we cover it in detail in that section.

5.4 *Developer Biases*

Diversity of developers is a problem of epic proportions especially when it comes to data-driven technologies. It explains all too many of the blunders we've seen in recent years, if we can call them that. In terms of binary gender thinking, approximately 80% of software developers are men: that's four-to-one (Cheryan et al., 2022; Klawe, 2020). If we narrow our pool to developers of data-driven technology, those numbers become worse. According to an AI Index survey, female faculty made up just 16.1% of all tenure track computer science faculty at several universities around the world in 2020 (AI Index Report, 2021). That year, only 15% of AI researchers at Facebook, and 10% of AI researchers at Google were women. Representation in the development of this technology is imperative, in the quest for inclusive technology.

Three antecedents to support this claim. The first is a data analysis experiment where 29 teams developed different solutions to the same problem related to bias (Silberzahn & Uhlmann, 2015). A second study showed that cognitive biases of developers were transferred to their code (Johansen et al., 2021). A third study showed that developer errors are correlated within communities (Cowgill et al., 2020). To put it simply, a more diverse set of voices catches more errors.

6 Biases in User Interaction

One significant source of bias comes from user interaction (not solely limited to the Web). These types of biases have two sources: the user interface and the biased interaction of the user or user bias. The first key bias in the user interface is called exposure or presentation bias: everything that is exposed to the user has a positive probability of being clicked, while everything else has none. This is particularly relevant for recommendation systems. Let us consider a video streaming service. Even if we have hundreds of recommendations that we can browse, that number is abysmally small compared to the millions of possibilities that might be out there. This bias will affect new items or items that have not previously been shown, since there is no usage data for them. The most common solution to this problem is called *explore and exploit* (see Agarwal et al. (2009) for a classic example applied to the Web. This technique exposes the user to new items to *explore*, randomly intermingled with top recommendations. The idea being that information from the (new) items chosen can be exploited to improve recommendations in the future. The paradox of this technique is that exploring may imply a loss, that is the opportunity cost of exploiting information already known. In some cases, there is even a revenue loss, such as in the case of digital ads. However, in the long term, as the system knows the market better, the revenue can be larger (Delnevo & Baeza-Yates, 2022). From the perspective of the user, the best recommendations will always be the things you wouldn't have otherwise known about.

The second relevant bias is position bias. For instance, in Western cultures, we read from top to bottom and from left to right. Our bias is to look first toward the top left corner of the screen prompting that region of the screen to get more clicks. An important instance of this bias is ranking bias. Consider a web search engine where result pages are listed in relevant order from top to bottom. The top ranked result will get more clicks than the others because it is both the (probably) most relevant result but also is in the first position. To be able to use click data for improving and evaluating ranking algorithms, we must debias the click distribution; otherwise, feedback in our algorithms will simply amplify already popular pages.

Other biases in the user interaction include additional effects of user interaction design. For instance, any content you need to scroll to see will suffer from exposure bias. Content near images will have a larger probability of being clicked because images attract our attention. Examples from eye-tracking studies show that since *universal search*[3] was introduced, the non-text content counteracts ranking bias in the results (Mediative, 2014).

Social bias defines how other peoples' content affects our judgment. One example comes from collaborative ratings: assume you want to rate an item with a low score, and you see that most people have a high score. You may increase your score assuming that perhaps you are being harsh. This bias has been already explored for Amazon reviews data (Wang & Wang, 2014, pp. 196–204) and may also be referred to as social conformity or the herding effect (Olteanu et al., 2016).

Finally, the way that each person interacts with any type of device is very personal. Some people are eager to click, while other people move the mouse to where one is looking. Mouse movement is a partial proxy for gaze attention and, in turn, a cheap replacement for eye-tracking. Some people may not notice the scrolling bar, or some people like to read in detail while others just skim. In addition to the bias introduced by interaction designers, we have our own cultural and cognitive biases. A good example of how cultural and cognitive biases affect web search engines is presented by White (2013), where it is shown that users tend to select results aligned with their beliefs, or confirmation bias. To make the problem even more complex, interaction biases cascade in the system and isolating each one is difficult. In Fig. 7, we show an example of how these biases cascade and depend on each other, implying that we are always seeing their composed effects. For instance, ranking bias is an instance of position bias as users tend to click in top results. Similarly, users that scroll affect how they move the mouse as well as which elements of the screen they can click.

The interaction biases just explained are crucial as many web systems are optimized by using implicit user feedback. As those systems are usually machine learning based, they learn to reinforce their own biases or the biases of linked systems, yielding suboptimal solutions and/or self-fulfilling prophecies. Sometimes these systems even compete among themselves, such that an improvement in one system results from a degradation in another system that uses a different (inversely

[3] Universal search results include other media in addition to text, such as images and videos.

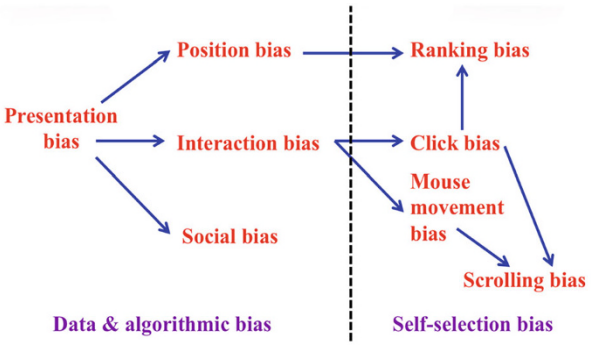


Fig. 7 Dependency graph of some biases that affect the user interaction

correlated) optimization function. A classic example of this is the tension between user experience and monetization teams in Internet companies.

7 The Vicious Cycle of Bias

Bias begets bias. Imagine that you are a blogger planning your next entry. First, you search for pages about the topic you wish to cover. Second, you select a few sources that seem relevant to you. Third, you select some quotes from those sources. Fourth, you write the new content, putting the quotes in the right places, of course citing the sources. Finally, you happily publish the new entry on the Web.

The content creation process outlined does not apply solely to bloggers but also to content in reviews, comments, posts, tweets, toots, and more. The problem occurs when a subset of results is returned, based on what the search engine encodes as relevant to the query. In this way, the ranking algorithm creates a feedback loop, simply because the content that is shortlisted, gets duplicated, and amplified over time. In a study that we did a few years ago, we found that about 35% of the content of the Chilean Web was duplicated and we could trace the genealogy of the partial (semantic) duplication of those pages (Baeza-Yates et al., 2008, pp. 367–376). Today, this effect probably is much larger.

The process above creates a vicious cycle of feedback loop bias because some content providers get more link references which lead to more clicks. Even if you debias them, the rich get richer. Furthermore, the duplication of content makes the problem of distinguishing good pages from bad more complex. Web spammers in turn reuse content from good pages to fake quality content, which only adds to the problem. So paradoxically, search engines are harming themselves unless they do not account for all the biases involved.

Another example of feedback loop bias comes from personalization algorithms, or what Eli Pariser describes as the filter bubble (Pariser, 2011). Personalization of course means that different people making the same query need not see the same results. The argument for personalization is clear: humans need help both filtering

Table 1 Our proposed classification of biases

Bias/type	Statistical	Cultural	Cognitive
Algorithmic	◆	◆	◆
Exposure	◆		
Position	◆		
Developer		◆	◆
Data	◆	◆	◆
Sampling	◆		
Linguistic		◆	
Visual		◆	
Feedback	◆	◆	◆
Engagement		◆	◆
User interaction		◆	◆
Ranking		◆	◆
Social		◆	◆
User		◆	◆

and finding information. But personalization algorithms can also shape our perception of the world. For instance, take an algorithm that relies on our interaction data to show us things we'd "like," filtering out less likable content that is important on some other dimensions not deemed of no advantage to the creators of the technology. At macrolevel, this technology poses the risk of creating social echo chambers that misinform at the behest of foreign or private interests, hindering collective social progress. This issue must be counteracted with collaborative filtering or task contextualization as well as promoting diversity, novelty, serendipity, and even exposure to counterarguments. Such strategies also have a positive impact on privacy online because solutions incorporating them require less personal information.

8 Conclusions

The problem of bias is far more complex than outlined here. We cover just part of the Web, the tip of the bias iceberg so to speak. At its foundation reside our individual and collective biases. On the contrary, many of the biases described here are valid beyond the Web ecosystem, through mobile devices and the Internet of Things.

In Table 1, we attempt to classify the biases described above, as statistical, cultural, or cognitive, by marking the appropriate column. Some instances are a combination of all three. At the top of the table are pure algorithmic biases, though as we've seen, each program inevitably encodes the cultural and cognitive biases of their creators. The lower group includes those biases arising from people while the middle group includes biases where algorithms are involved.

In October 2022, ACM published their second statement on principles for responsible algorithmic systems (ACM Tech Policy Council, 2022). These are legitimacy and competency; minimizing harm; security and privacy; transparency;

interpretability and explainability; maintainability; contestability and auditability; accountability and responsibility; and limiting environmental impacts. The goal of this article is aligned with several principles including minimizing harm (bias) and transparency (bias awareness). In addition, at least two new conferences that address this topic were started in 2018, FAccT and AIES. All these efforts should help our community as we define algorithmic ethics, particularly with respect to machine learning.

Finally, any attempt to be unbiased might be already biased with our own cultural and cognitive biases. The first step is to be aware of all these biases. Only by knowing of their existence can we hope to grapple with and mitigate them. The alternative is a world without fact, where decisions are made based on biased perceptions, in which no amount of diversity, novelty, or serendipity can save us.

Discussion Questions for Students and Their Teachers

1. Discuss possible cognitive biases that may impact the Web and are not mentioned in this chapter. Finding a good taxonomy of cognitive biases is a good way to start.
2. Name all sources of bias that you can think of and discuss how they are related. Mapping the examples of this chapter as well as others to the sources helps.
3. An example of non-trivial reference value is how many web pages in a language should be. What is the right value to measure for bias? Who should decide that?
4. If the bias of the developers is transferred to their code, should developing teams be more diverse? Or are there cases where we may want certain demographics in the team such that the best possible system is built?
5. Assume that you find two different biases that are positively correlated. How can you decide if one of them causes the other or that they are independent?

Learning Resources for Students

1. Persuading programmers to detect and mitigate bias in technology design: The role of motivational appeals and the speaker (Almánzar et al., 2023)

   This paper proposes and studies a conceptual framework for the effectiveness of motivational appeals aimed at programmers, considering the role of framing, the speaker's race and gender, and the individual differences in recipients' social dominance orientation egalitarianism (SDO-E) in driving bias detection outcomes. They suggest that a problem framing, "You are part of the problem," will be more effective than a solution framing, "You are part of the solution," when the speaker is White and male rather than Black and female, but this only applies to respondents with low levels of SDO-E and will be reversed for respondents with high levels of SDO-E, due to the pursuit of egalitarian values that automatically inhibits the activation of stereotypes.
2. Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers (Peng et al., 2021).

   This paper analyzes three influential but problematic datasets on face recognition that are used by almost 1000 papers. They find that derivative datasets and models, broader technological and social change, the lack of clarity of licenses, and dataset management practices can introduce additional ethical issues,

proposing a distributed approach to harm mitigation that considers the full life cycle of a dataset.

3. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries (Olteanu et al., 2019)

   This paper surveys several issues of social data: (1) biases and inaccuracies occurring at the source of the data, but also introduced during processing; (2) methodological limitations and pitfalls; and (3) ethical boundaries and unexpected consequences that are often overlooked. As a result, they present a framework for identifying a broad variety of dangers in the research and practices around social data use.

4. Taxonomy of Risks posed by Language Models (Weidinger et al., 2022)

   This paper categorizes language model risks into six broad subgroups, some of which have been touched on in this chapter. A more complete picture is provided by the referenced publication. One area not discussed here are those around "human-computer interactions." As machines become more competent at emulating ever increasing modes of human communication, what might be the benefits and risks of such technology?

5. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (Schwartz et al., 2022)

   This document covers the challenging area of AI bias, providing a first step on the roadmap for developing detailed sociotechnical guidance for identifying and managing AI bias. Specifically, they (1) describe the stakes of bias in AI intelligence and provides examples of how and why it can chip away at public trust; (2) identify three categories of bias in AI—systemic, statistical, and human—and describe how and where they contribute to harms; and (3) describe three broad challenges for mitigating bias, namely, datasets, testing and evaluation, and human factors, and recommendations for addressing them.

6. Ethical Development (Murgai, 2023)

   Chapter 2 of the referenced resource discusses how to go about developing machine learning applications ethically. It focuses on practical aspects of developing a data and model governance framework and provides a taxonomy of common causes of harm relating them to the stage of the workflow at which they can be detected and prevented.

References

ACM Tech Policy Council. (2022). *Statement on responsible algorithmic systems. 26 October 2022*. https://www.acm.org/binaries/content/assets/public-policy/final-joint-ai-statement-update.pdf

Agarwal, D., Chen, B.-C., & Elango, P. (2009). Explore/exploit schemes for web content optimization. In *Proceedings of the Ninth IEEE International Conference on Data Mining*. IEEE Computer Society.

Almánzar, A. R., Edinger-Schons, L. M., & Grüning, D. J. (2023). Persuading programmers to detect and mitigate bias in technology design: The role of motivational appeals and the speaker. *PsyArXiv*. https://doi.org/10.31234/osf.io/jbxeq

Artificial Intelligence Index Report. (2021). *Diversity in AI*. https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report-_Chapter-6.pdf.

Baeza-Yates, R. (2015). *Incremental sampling of query logs*. Industry track. In *Proceedings of the 38th ACM SIGIR Conference* (pp. 1093–1096).

Baeza-Yates, R. (2018). Bias on the Web. *Communications of the ACM*, June, *61*(6), 54–61. https://doi.org/10.1145/3209581.

Baeza-Yates, R., & Castillo, C. (2006). *Relationship between web links and trade* (poster). In Proceedings of the 15th international conference on the World Wide Web (pp. 927–928).

Baeza-Yates, R., & Saez-Trumper, D. (2015). Wisdom of the crowd or wisdom of a few? An analysis of users' content generation. In *Proceedings of the 26th ACM conference on hypertext and social media* (pp. 69–74).

Baeza-Yates, R., Castillo, C., & López, V. (2006). Characteristics of the Web of Spain. *El Profesional de la Información, 15*(1), 1–17.

Baeza-Yates, R., Castillo, C., & Efthimiadis, E. N. (2007). Characterization of national web domains. *ACM Transactions on Internet Technology, 7*(2).

Baeza-Yates, R., Pereira, Á., & Ziviani, N. (2008). Genealogical trees on the Web: A search engine user perspective. In *Proceedings of the 17th international conference on the World Wide Web* (pp. 367–376).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the ACM conference on fairness, accountability, and transparency (FAccT'21)*. Association for Computing Machinery, , 610–623. doi:https://doi.org/10.1145/3442188.3445922.

Bhutada, G. (2021). *Visualizing the most used languages on the Internet*. March 26. https://www.visualcapitalist.com/the-most-used-languages-on-the-internet/

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th conference on neural information processing systems*.

Boroditsky, L. (2017). *How language shapes the way we think*. TEDWomen. https://www.ted.com/talks/lera_boroditsky_how_language_shapes_the_way_we_think

Boroditsky, L., & Gaby, A. (2010). Remembrances of times east: Absolute spatial representations of time in an Australian aboriginal community. *Psychological Science, 21*(11), 1635–1639. https://doi.org/10.1177/0956797610386621

Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. *Language in Mind: Advances in the Study of Language and Thought, 22*, 61–79.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. vol. 81. In *Proceedings of Machine Learning Research* (pp. 1–15).

Cadwalladr, C. (2019). *Facebook's role in Brexit -- and the threat to democracy*. TED. https://www.ted.com/talks/carole_cadwalladr_facebook_s_role_in_brexit_and_the_threat_to_democracy.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.

Carylsue. (2016). *New Guinea natives navigate by valleys and mountains*. National Geographic Education Blog, Geography. April 14. https://blog.education.nationalgeographic.org/2016/04/14/new-guinea-natives-navigate-their-homes-by-valleys-and-mountains/.

Cheryan, S., Master, A., & Meltzoff, A. (2022). There are too few women in computer science and engineering. *Scientific American*, July 27. https://www.scientificamerican.com/article/there-are-too-few-women-in-computer-science-and-engineering/.

Chouldechova, A. (2017). *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.* doi:https://doi.org/10.1089/big.2016.0047.

Cima, R. (2015). How photography was optimized for white skin color. *Priceonomics*. https://priceonomics.com/how-photography-was-optimized-for-white-skin/.

Corn Jr, R. (M). (2019). Native American culture - Language: the key to everything. *TEDxOshkosh*. January 24. https://www.ted.com/talks/ron_muqsahkwat_corn_jr_native_american_culture_language_the_key_to_everything

Costa-jussà, M. R., Lin, P. L., & España-Bonet, C. (2020). GeBioToolkit: automatic extraction of gender-balanced multilingual corpus of Wikipedia biographies. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 4081–4088). European Language Resources Association, Marseille.

Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N., & Chaintreau, A. (2020). Biased programmers? Or biased data? *A field experiment in operationalizing AI ethics*. https://doi.org/10.48550/arXiv.2012.02394.

Crawford, K. (2017). *The trouble with bias*. NIPS Keynote.

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data, 5*(2), 120–134.

Del Barco, M. (2014). *How Kodak's Shirley cards set photography's skin-tone standard*. NPR KQED, November 13. https://www.npr.org/2014/11/13/363517842/for-decades-kodak-s-shirley-cards-set-photography-s-skin-tone-standard

Delnevo, G., & Baeza-Yates, R. (2022). *Exploration trade-offs in web recommender systems*. IEEE Big Data.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, & I. Wegener (Eds.), *Automata, languages and programming. ICALP 2006* (Lecture notes in computer science) (Vol. 4052). Springer. https://doi.org/10.1007/11787006_1

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference (ITCS '12)* (pp. 214–226). Association for Computing Machinery. https://doi.org/10.1145/2090236.2090255

Dyer, R. (1997). *White*. Routledge.

Fausey, C. M., & Boroditsky, L. (2008). English and Spanish speakers remember causal agents differently. *Proceedings of the Annual Meeting of the Cognitive Science Society, 30*(30).

Fetterly, D., Manasse, M., & Najork, M. (2003) On the evolution of clusters of near-duplicate web pages. *Proceedings of the IEEE/LEOS 3rd international conference on numerical simulation of semiconductor optoelectronic devices (IEEE Cat. No. 03EX726)*, Santiago, pp. 37–45, doi: https://doi.org/10.1109/LAWEB.2003.1250280.

Gong, W., Lim, E-P, & Zhu, F. (2015). Characterizing silent users in social media communities. In *Proceedings of ninth international AAAI conference on Web and Social Media*.

Graells-Garrido, E. & Lalmas, M. (2014). Balancing diversity to counter-measure geographical centralization in microblogging platforms. In *Proceedings of the 25th ACM conference on hypertext and social media* (pp. 231–236).

Graells-Garrido, E., Lalmas, M., & Menczer, F. (2015). First women, second sex: Gender bias in Wikipedia. In *Proceedings of the 26th ACM conference on hypertext and social media* (pp. 165–174).

Hale, L., & Peralta, E. (2021). *Social media misinformation stokes a worsening civil war in Ethiopia*. NPR, October 15. https://www.npr.org/2021/10/15/1046106922/social-media-misinformation-stokes-a-worsening-civil-war-in-ethiopia

Hartsfield, T. (2019). *ChatGPT answers physics questions like a confused C student*. February. https://bigthink.com/the-present/chatgpt-physics/.

Johansen, J., Pedersen, T., & Johansen, C. (2021). Studying human-to-computer bias transference. *AI & Society*. https://doi.org/10.1007/s00146-021-01328-4

Kahneman, D. (2011). *Thinking, fast and slow*.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Machine learning and knowledge discovery in databases. ECML PKDD* (Lecture Notes in Computer Science) (Vol. 7524). Springer. https://doi.org/10.1007/978-3-642-33486-3_3

Kay, M., Matuszek, C., & Munson, S. A. (2015). *Unequal representation and gender stereotypes in image search results for occupations*. ACM.

Kearns, M., & Roth, A. (2019). *The ethical algorithm*. Talks at Google. https://youtu.be/tmC9JdKc3sA.

Klawe, M. (2020). *Why diversity in AI is so important*. July 16. https://www.forbes.com/sites/mariaklawe/2020/07/16/why-diversity-in-ai-is-so-important/

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). *Human decisions and machine predictions*. https://www.nber.org/papers/w23180.

Kuczmarski, J. (2018). *Reducing gender bias in Google Translate*. Google blog, December 6. https://blog.google/products/translate/reducing-gender-bias-google-translate/.

Lazer, D. M. J., et al. (2018). The Science of Fake News. *Science, 359*(6380), 1094–1096.

Lazovich, T., Belli, L., Gonzales, A., Bower, A., Tantipongpipat, U., Lum, K., Huszar, F., & Chowdhury, R. (2022). Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics. *Patterns, 3*(8).

Mediative. (2014). *The evolution of Google's search results pages and effects on user behavior* (white paper). http://www.mediative.com/SER.

Murgai, L. (2023). *Mitigating bias in machine learning*. https://mitigatingbias.ml.

Naranjo, J. (2023). *Ethiopia's forgotten war is the deadliest of the 21st century, with around 600,000 civilian deaths*. El País. Jan 27, 2023. https://english.elpais.com/international/2023-01-27/ethiopias-forgotten-war-is-the-deadliest-of-the-21st-century-with-around-600000-civilian-deaths.html.

Nielsen, J. (2006). *The 90-9-1 rule for participation inequality in social media and online communities*. October 8. https://www.nngroup.com/articles/participation-inequality/.

O'Reiley, L. (2017). *Facebook's claimed reach in the U.S. is larger than census figures, analyst finds*. WSJ Sept. 6. https://www.wsj.com/articles/facebooks-claimed-reach-in-the-u-s-is-larger-than-census-figures-analyst-finds-1504711935

O'Toole, E. (2016). A dictionary entry citing 'rabid feminist' doesn't just reflect prejudice, it reinforces it. *The Guardian*, January 26. https://www.theguardian.com/commentisfree/2016/jan/26/rabid-feminist-language-oxford-english-dictionary

Olteanu, A., Castillo, C., Diaz, C., & Kiciman, E. (2016). *Social data: Biases, methodological pitfalls, and ethical boundaries*. Available at SSRN: https://ssrn.com/abstract=2886526.

Olteanu, A., Carlos Castillo, C., Fernando Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data, 2*. https://doi.org/10.3389/fdata.2019.00013

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin.

Peng, K., Mathur A., & Narayanan, A. (2021). Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Proceedings of the neural information processing systems track on datasets and benchmarks*.

Perez, C. C. (2019). *Invisible women: Data bias in a world designed for men*. Vintage Books.

Pitel, L. (2019). *Can a genderless language change the way we think*. FT, August 5. https://www.ft.com/content/7b59352c-b75e-11e9-8a88-aa6628ac896c.

Prabhu, V. U., & Birhane, A. (2021). *Large image datasets: A pyrrhic win for computer vision?* Available: https://arxiv.org/abs/2006.16923

Rose, A. (2010). Are face-detection cameras racist? *Time*. January 22. https://content.time.com/time/business/article/0,8599,1954643,00.html.

Saez-Trumper, D., Castillo, C., & Lalmas, M. (2013). Social media news communities: Gatekeeping, coverage, and statement bias. In *ACM CIKM* (pp. 1679–1684).

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence*. National Institute of Standards and

Technology Special Publication 1270, USA. Freely available at doi:https://doi.org/10.6028/NIST.SP.1270.
Shariatmadari, D. (2016). Eight words that reveal the sexism at the heart of the English language. *The Guardian*, January 27. https://www.theguardian.com/commentisfree/2016/jan/27/eight-words-sexism-heart-english-language.
Silberzahn, R., & Uhlmann, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature, 526*, 189–191, October 2015. Full report is available at https://psyarxiv.com/qkwst/, 2017.
Smith, M., Patil, D. J., & Muñoz, C. (2016). *Big data: A report on algorithmic systems, opportunity, and civil rights*. Executive Office of the President.
Spice, B. (2015). *Fewer women than men are shown online ads related to high-paying jobs*. CMU CSD. July. https://csd.cmu.edu/news/fewer-women-men-are-shown-online-ads-related-high-paying-jobs.
Statista. (2021). *Languages on the Internet*. https://www.statista.com/chart/26884/languages-on-the-internet/.
Sweeney, L. (2013). Discrimination in online ad delivery. SSRN. https://ssrn.com/abstract=2208240.
Trask, L. (2007). *Trask's historical linguistics*. Routledge.
Truong, K. (2020). *This image of a White Barack Obama is AI's racial bias problem in a Nutshell*. June. https://www.vice.com/en/article/7kpxyy/this-image-of-a-white-barack-obama-is-ais-racial-bias-problem-in-a-nutshell.
Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *AAAI ICWSM Conference*, pp. 454–463.
Wang, T., & Wang, D. (2014). Why Amazon's ratings might mislead you: The story of herding effects. *Big Data, 2*(4), 196–204.
Wattles, J. (2015). *Amazon sues more than 1,000 sellers of 'fake' product reviews*. October 19. https://money.cnn.com/2015/10/18/technology/amazon-lawsuit-fake-reviews/index.html.
Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. & Gabriel, I. (2021). *Ethical and social risks of harm from Language Models*.
Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L.A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., & Gabriel, I. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency (FAccT '22)* (pp. 214–229). Association for Computing Machinery. doi:https://doi.org/10.1145/3531146.3533088.
Weinberg, J. (2016). *Cognitive bias codex*. https://dailynous.com/2016/09/14/cognitive-bias-codex/.
White, R. (2013). Beliefs and biases in web search. In *Proceedings of the 36th ACM SIGIR conference*, pp. 3–12.
Winawer, J., Witthoft, N., Frank, M. C., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proc Natl Acad Sci U S A, 104*(19), 7780–7785. https://doi.org/10.1073/pnas.0701644104
Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on Twitter. In *Proceedings of the 20th international conference on the World Wide Web* (pp. 705–714). ACM Press.
Young, V. A. (2020). Nearly half of the Twitter accounts discussing 'Reopening America' may be bots. *CMU News*, May 27. https://www.cmu.edu/news/stories/archives/2020/may/twitter-bot-campaign.html.
Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.