Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–June: 1–14
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517231155060
journals.sagepub.com/home/bds
SAGE

Because the machine can discriminate: How machine learning serves and transforms biological explanations of human difference

Jeffrey W. Lockhart[1]

Abstract

Research on scientific/intellectual movements, and social movements generally, tends to focus on resources and conditions outside the substance of the movements, such as funding and publication opportunities or the prestige and networks of movement actors. Drawing on Pinch's theory of technologies as institutions, I argue that research methods can also serve as resources for scientific movements by institutionalizing their ideas in research practice. I demonstrate the argument with the case of neuroscience, where the adoption of machine learning changed how scientists think about measurement and modeling of group difference. This provided an opportunity for members of the sex difference movement by offering a 'truly categorical' quantitative methodology that aligned more closely with their understanding of male and female brains and bodies as categorically distinct. The result was a flurry of publications and symbiotic relationships with other researchers that rescued a scientific movement which had been growing increasingly untenable under the prior methodological regime of univariate, frequentist analyses. I call for increased sociological attention to the inner workings of technologies that we typically black box in light of their potential consequences for the social world. I also suggest that machine learning in particular might have wide-reaching implications for how we conceive of human groups beyond sex, including race, sexuality, criminality, and political position, where scientists are just beginning to adopt its methods.

Keywords

Gender, sex, machine learning, neuroscience, scientific/intellectual movements, research methods

Scientists have long engaged in a set of related debates about whether human social groups of race, class, gender, and more constitute biologically distinct categories, and these debates remain lively today (Lockhart, 2021; Morning, 2014; Sanz, 2017). I argue that part of the way they have remained lively in recent decades is through the adoption, on all sides, of machine learning (ML). In this paper, I focus on how the introduction of ML to neuroscience rescued the foundering scientific/intellectual movement (SIM) for sex differences, even elevating its object to a gold standard, by preserving core theories and agendas while radically altering the field's quantitative reasoning.

In particular, drawing on theories of technologies as institutions (Pinch, 2008), I argue that ML serves to institutionalize a frame of understanding for group difference that is favorable to the sex difference movement as part of general research practice in neuroscience. This means that ML and research methods in general can serve as resources for SIMs, extending existing theorizations that focus on traditional factors such as funding, publication, and prestige (Frickel and Gross, 2005) to show that the content and practice of research itself can be key resources for movements. Understanding this process requires not only a close analysis of how the use and results of a method are framed by SIM members but also an analysis of the implementation and workings of the method itself as a technology, to understand how SIM norms and values are delegated to and materialized in it (Bucher, 2016).

While the values and frames embedded in ML have been extensively studied (e.g. Benjamin, 2019; Chun, 2021; Eubanks, 2017; Fourcade and Healy, 2016; Keyes, 2018; Noble, 2018; Scheuerman et al., 2021; Seaver, 2018), this

[1] Department of Sociology, University of Chicago, Chicago, IL, USA

Corresponding author:
Jeffrey W. Lockhart, Department of Sociology, University of Chicago, 1155 East 60th Street, Chicago, IL 60637, USA.
Email: jlockhart@uchicago.edu

work focuses almost exclusively on production algorithms (those used by governments and companies to *do* things like set prices, recommend songs, deny parole, or target missiles). Little attention has gone to the ways values and frames are embedded in discovery algorithms (ML used by scientists to *learn* things, make claims about the world, or advance SIMs) (Keyes et al., 2021), even though sociologists have long known that research methods are 'device[s] for transforming common-sense meanings and implicit theoretical notions into acceptable "evidence"' (Cicourel, 1964, p. 21).

I demonstrate the role discovery algorithms can play in institutionalizing SIM frames using the case of neuroscientific research on sex, where competing SIMs debate whether men's and women's brains are categorically distinct (Epstein, 2007; Lockhart, 2020; Sanz, 2017). ML is the latest in a long parade of new research tools and methods adopted to study sex differences in brains after previous approaches became untenable (Lockhart, 2020). Like those methods, ML exhibits interpretive flexibility: scientists using the same data, methods, and results can come to a variety of conflicting conclusions (Collins, 1992). ML is used to argue both for and against a categorical distinction between men's and women's brains, although the movement for categorical differences is much larger than its countermovement.

But ML is unlike prior innovations, which focused on things like finding sex in new parts of the brain or normalizing brain sizes in new ways. Instead, ML represents a dramatic shift in quantitative reasoning about group differences. As I show, ML's classification approach is far better aligned with the core question in the debate (viz. 'are men and women categorically different?') than previously dominant statistical approaches. Additionally, ML comes with a thriving community of computer science researchers who, rather than being tied to a particular substantive domain, seek out classification tasks in order to innovate, optimize, and compete with one another. For computer scientists working with neuroimaging data, sex difference became the gold standard they competed to maximize rather than an empirical question to be tested.

Sex science is particularly important today, as substantial social and political conflicts at the local, national, and international levels turn on interpretations of scientific claims about the biology of sex (Sudai et al., 2022). Moreover, despite considerable attention to the role of ML in government and corporate contexts, comparatively little scholarship has investigated the ways ML shapes and enables work in scientific contexts (Keyes et al., 2021).

Background

SIMs are 'collective efforts to pursue research programs or projects for thought in the face of resistance from others in the scientific or intellectual community' (Frickel and Gross, 2005, p. 208). They take a variety of forms, challenging or defending the status quo, advancing progressive directions, or reactionarily reviving past ideas. And like other movements, their success depends in part on structural conditions and access to resources. Frickel and Gross enumerate some of the obvious resources for scientists: funding, publication opportunities, and employment are all key to the long-term viability of a research program. But as *intellectual* movements, the ability to frame their ideas in ways that resonate with scientists in the target field(s) is critical: the goal is to shape the field of ideas. Frickel and Gross focus on the importance of framing a SIM's motivating problem to get buy-in from other scientists, but a SIM's proposed solution/research program must also be successfully framed. And this, I argue, is where research methods can serve as resources for SIMs by institutionalizing their frames.

Here I turn to Pinch's (2008, p. 467) theorization of technologies as institutions, 'sets of rules or patterns whereby social actions and practices are ordered' and made reproducible. By this definition, research methods are clearly institutions, ordering the practices and actions of scientists in pursuit of reproducible results. Research methods are more than just the explicit guidance written into methods sections. They have affordances, things that they enable users and other technologies to do, and ways of doing those things. Pinch uses the example of the operating system, DOS. Because DOS is an important technology at the heart of Microsoft computers, many other technologies and user behaviors are structured to interface with it. When DOS was embedded in Windows, it receded from view. But the technology remained there, structuring how myriad other technologies are designed, how they operate, and how people use and think about computing. As Pinch concludes:

> The embedding or freezing of choices within scientific and technical systems, what the French philosopher Gaston Bachelard calls *phenomenotechnique*, makes technology actually one of the most powerful institutions in Jepperson's sense we as social scientists face. It is because social choices appear to have vanished from technologies, or are so deeply embedded within technical structures that they become invisible to all but the technical experts, that technologies are powerful institutions (Pinch, 2008, p. 467).

The particular technologies at play in my case are classifiers, research methods from statistics, and ML used to sort people into categories of male and female. Their predictive approach is fundamentally different from the frequentist inferential statistics most scientists are used to, involving different workflows and reasoning (Salganik, 2017). Krippner and Hirschman (2022) show how the frames institutionalized by two older classification technologies—class-based insurance pricing and attribute-based credit scoring—can enable or disable social movements. The class-based insurance pricing system assigned people to groups according to shared attributes and then charged one rate for all members of each group. This allowed the women's movement to mobilize around price discrimination: groups of women and men were salient in both the movement and the technology, and they were charged different rates. Attribute-based credit scoring, in contrast, is more like regression. People are 'grouped' by score, but critically, two people with the same score might have entirely different attributes, and the system is built to individualize each person into their own score rather than establishing a group of similar people. The result was demobilizing: 'attribute-based systems of classification, which not only suppress but also scramble group identities, mak[e] collective mobilization considerably more difficult' (Krippner and Hirschman, 2022, p. 21).

The increased complexity of attribute-based scoring and modern ML compared with other classification approaches makes it more opaque, which in turn risks granting the classifications they construct the 'naturalized facticity characteristic of classic social facts' (Fourcade and Healy, 2017, p. 286). Indeed, humans' *inability* to fully grasp the complex patterns that ML relies on can lead to a sort of 'machinic neoplatonism' or the belief that ML has revealed a hidden mathematical truth that is beyond human questioning or verification (McQuillan, 2018).

Thus, I argue that if a SIM can find a technology that embeds its preferred idea and use it as a research method, that method can institutionalize the idea, letting it recede from view while substantially influencing the mobilization possibilities for both the SIM and its opponents. In the case of the scientific movement for sex difference research (described in Epstein, 2007; Lockhart, 2020), I argue that ML classifiers, especially support vector machines (SVM), institutionalize the movement's frame of human difference as constituted by categorically separate groups.

Scientists have long measured skulls and brains to distinguish human groups. In his infamous 1879 essay on the subject, Gustave Le Bon argued that:

> In the most intelligent races,… there are a large number of women whose brains are closer in size to those of gorillas than to the most developed male brains. This inferiority is so obvious that no one can contest it for a moment; only its degree is worth discussion… there exist some distinguished women, very superior to the average man, but they are as exceptional as the birth of any monstrosity, as for example of a gorilla with two heads; consequently, we may neglect them entirely (translated in Gould, 1996, pp. 136–7).

While scientists no longer pack skulls with lead shot to measure cranial volume or assume brain size corresponds directly with intelligence, they do continue to publish comparisons of men's and women's brains, often sharing two of Le Bon's core assumptions. First, Le Bon declares the difference between men and women obvious, thereby narrowing the scope of legitimate scientific inquiry to merely the measurement of difference, not the fact of difference. Second, he acknowledges that differences are not truly categorical but asserts that scientists should neglect exceptions, thereby ensuring conclusions are in line with the assumption of categorical difference.

Biological sex, like gender, is socially and scientifically constructed. Which material parts of bodies count as sex, how they relate to each other, the causes of those relationships, and their meanings are ever-changing and contested (Fausto-Sterling, 2005; Laqueur, 1990; Lockhart, 2020). Claims that brains are sites of sex differences have always met resistance from a countermovement of scientists opposed to them. These scientists recenter the question of whether human brains should be thought of as coming in two distinct forms, male and female, often emphasizing the exceptions Le Bon insisted on neglecting (Lockhart, 2020; Sanz, 2017). Scientists on both sides tend to view the debate in terms of group conflict and struggle for control over research on sex, with the most active participants routinely publishing lists of 'way[s] to defeat the idea[s]' of 'anti-sex difference authors' (Cahill, 2014, p. 8) or lamenting that they must endlessly 'dispatch' the '"Whac-a-Mole" myths' of sex differences (Rippon, 2019, p. xii).

A key battleground for these SIMs is neuroscience, which began to adopt ML in the early 2000s. The field's phrenological baggage was catching up to it, and a combination of new technologies and research failures made older statistical methodologies untenable. Phrenology and related sciences treated the brain as a collection of organs, whose size and shape directly corresponded with particular capacities (Anderson, 2014): One part of the brain for color sense, another for religiosity, and so on. This understanding co-developed with statistics, whose practitioners were often eugenically interested in establishing differences between human groups (Clayton, 2020). The statistical tools they developed allowed comparisons between individual variables: one brain region or measurement could be correlated with one behavior, or one measure could be compared in two groups. This is known as univariate analysis. Even later developments like multiple regression are still largely thought of as tools to evaluate a single variable (e.g. after 'controlling for' other variables to remove their influence).

Technical advances put pressure on this framework. New measurement devices like MRI gave scientists access to datasets that had tens of thousands or even millions of variables, each representing, for example, one cubic millimeter of a brain. Neuroscientists adapted their univariate methodology into the 'mass univariate' approach. Much like geneticists were doing with genome-wide association studies, neuroscientists conducted vast numbers of separate statistical tests for a single paper, asking, ‘is this cubic millimeter of the brain correlated with sex? what about the one next to it?’ and so on for the whole brain. Mass univariate analysis is premised on an updated version of the phrenological assumption that neuroscientists today call ‘strong modularity hypothesis’, which holds that ‘every possible face, animal, and object category has a specialized [brain] region or set of neurons dedicated to its representation’ (Haxby, 2012, p. 853).

But scientists’ belief in the strong modularity hypothesis faltered. The dramatic results scientists hoped for never materialized with univariate and mass univariate approaches, and the smaller findings they published rarely survived replication. This led neuroscientists and geneticists to turn to ever more complex, multivariate ‘witches’ brew’ and ‘psychobiosocial’ theories involving vast networks of interactions and feedback cycles that simply cannot be modeled with the univariate statistical tools they had available (Arribas-Ayllon et al., 2010; Wade, 2013). The strong modularity hypothesis ‘didn’t seem possible. There are too many ways that faces and objects can be categorized’ for each to have a dedicated part of the brain (Haxby, 2012, p. 2). Neuroscientists still believed the things they studied existed in brains, but they seemed not to exist in the way scientists had imagined and searched for them.

ML offered a way out. ML entered neuroimaging under the name ‘multivariate pattern analysis’ in an attempt to match methods with the revised neuroscientific theories. Those theories still held that groups and phenotypes (like ‘male’ or ‘aggressive’) exist in brains, but no longer held they could be localized to isolated attributes (Haxby, 2012). Scientists argued that ML’s ability to find diverse and complicated relationships among variables offered a way to finally escape the phrenological assumption of a direct correspondence between brain regions and outcomes (Anderson, 2014). If identities, behaviors, abilities, and mental states were not individual areas but sets of conditions or constellations of interactions, then the statistical apparatus of ML was better suited to finding them. Proponents’ interest in developing a ‘fully postphrenological neuroscience’ using ML is entirely about moving on from an unfruitful research paradigm (Anderson, 2014). The book-length treatise *After Phrenology,* for example, does not use the words ‘race’, ‘racism’, ‘sex’, or ‘gender’ at all, and it opens by declaring ‘phrenology wasn’t such a bad idea. Certainly it is not deserving of the degree of scorn that is heaped on it’ (Anderson, 2014, p. xiii).

As I will argue in the remainder of this paper, the introduction of ML to neuroscience provided an opportunity for the sex difference SIM (and its feminist countermovement) by changing the understanding of human groups or classes embedded in the field’s research methods.

Data and methods

To understand the role of ML in SIMs, I follow the sociology of science tradition and examine the artifacts of ML’s production and contestation, the traces left when it is made to account. Those traces are abundant in the documents produced in scientists’ laboratories, those ‘that detail and lay out technical specifications [such as] conference papers on machine learning’ (Bucher, 2016, p. 87; Fitsch, 2021; Latour, 2005). I collected every publication that met the following criteria: written in English[1], uses neuroimaging data (e.g. MRI), and uses modeling (broadly defined) to predict sex/gender. A recent review of brain sex differences research lists 12 such studies (Eliot et al., 2021). I used those as a starting point, reading them for citations and keywords to search in both my university library’s databases and Google Scholar. I checked each result for the inclusion criteria and saved the relevant papers. Throughout the process, I added to and refined my set of search keywords until the results were exhaustive (Schilt and Westbrook, 2009). I further gathered comments, replies, and retractions for each of the publications, as well as their supplemental information and code.

During my initial open coding (Emerson et al., 2011), I reviewed each paper’s citations and added four additional papers that I missed while searching. I also removed a few papers that turned out not to meet the inclusion criteria, including five studies measuring skull bones in order to identify human remains. The resulting data set consists of 53 papers (48 published and 5 preprints) through early 2021. For one software package used in a number of papers, libsvm, I accessed both the current code and historic versions of it back to 2012 on github in order to verify methodological details that were unspecified in the publications (Figure 1).

I draw on a mix of techniques from content analysis, actor-network theory, and ethnography to analyze my

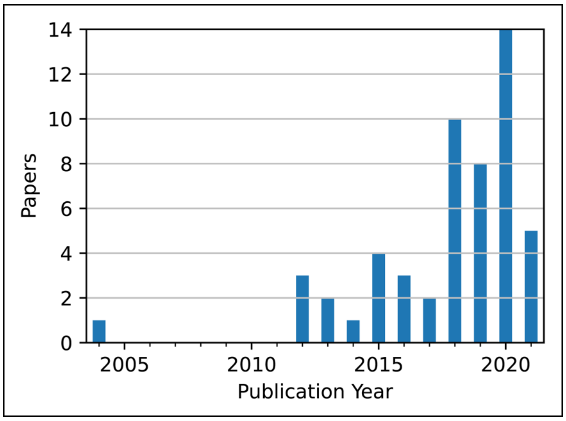

Figure 1. Papers using machine learning to predict sex from brain imaging data. Data for 2021 reflect only the first few months of that year.

data, a combination approach Bucher (2016) refers to as 'technography'. Technography observes technology to understand the interplay among people, with and through technology. It pays particular attention to how norms and values are delegated to and materialized in technology (Bucher, 2016, p. 86).

After my initial inductive, open coding of the data, I revisited all of the papers using a focused coding approach for 23 factors that emerged during open coding (Emerson et al., 2011), including technical details of imaging technologies, model types, accuracy, sample size, validation strategy, feature selection method, number of features, and disciplinary audience. During this second pass, I also gathered and coded every reference to ethics, research motivations, etiology of sex differences, how sex was measured, substantive interpretations of the findings, use of the word 'discrimination', and references to critics of categorical brain sex difference research.

Results

The 53 papers fall into several groups based on their stated goals, outlined in Table 1. Nearly half (26) of the papers are what I call sexual dimorphism papers—those that set out primarily to advance the SIM's cause of showing that human brains come in two categorically distinct forms, male and female. Almost a third (16) of the papers are proof-of-concept papers—those that set out to demonstrate a new method and used sex classification as their example. The authors of these papers have a symbiotic relationship with the sex difference SIM; using its ideas to illustrate their own work. Four papers are anti-essentialist—papers that set out to advance the countermovement by showing brains are not best described as a binary male/female categorization. A further three papers investigate whether transgender identity can be detected in brains. The remaining four papers do sex prediction, but set out primarily to study autism, pain, or psychological disorders.

Table 1. Sex prediction papers grouped by analytic category.

Goal of Paper	*N* Papers
Sexual dimorphism	26
Proof of concept	16
Anti-essentialist	4
Classify trans people	3
Misc.	4

The sex difference SIM's use of ML

A kerfuffle in the pages of the *Proceedings of the National Academy of Sciences* (*PNAS*) illustrates how members of the sex difference SIM use ML to advance their position that men and women have categorically distinct brains. A group of feminist neuroscientists led by Daphna Joel published a paper in the December 2015 issue. In it, they used descriptive statistics to argue that while specific attributes of brains might be associated with men or women on average, none was exclusive to men or women and any individual person's brain was in fact a heterogeneous 'mosaic' of both masculine and feminine features, such that no brain was categorically male or female (Joel et al., 2015). The response was immediate. Eight days after the paper was published, a team of sex difference SIM scientists posted a nine-page rebuttal to a preprint server. In order to argue that brains really are sexually dimorphic (i.e. they come in two distinct forms, one male and the other female), the preprint used ML to predict participant sex from Joel's brain MRI data, resulting in accuracies 'ranged from 68.5% to 77.2%' (Del Giudice et al., 2015, p. 8). Because ML could predict sex more accurately than random guessing (50%), they argued, the algorithm must be picking up on some real difference between men's and women's brains. The preprint further suggests, and tests, differentiating monkey species by face shape as an analogy to differentiating human men's and women's brains.

Within four months, *PNAS* published a shortened version of their preprint along with two other replies to Joel's article. Both other replies use ML the same way. The author of one explains why ML classification is a useful method for his argument: 'a classifier can only achieve perfect classification if the data points are well separated (note the converse does not hold: the data may be well separated, even if a particular classifier is no better than random guessing)' (Rosenblatt, 2016, p. E6966). He believes the research method can help his cause by showing male and female brains are 'well separated' but cannot help his opponents' cause because it cannot offer proof that is not separate. He concludes that 'brains are indeed typically male or typically female', because ML can classify sex from brains more accurately than random guessing ('about 80%') (Rosenblatt, 2016, p. E6966). The third letter concedes 'that a strict dichotomy between male/female brains does not exist' (Chekroud et al. 2016). Nevertheless, its authors insist that human men and women's brains are indeed categorically distinct in the same way that *cats and dogs* are. Their ML results for classifying brains according to sex show 'an individual's biological sex can be classified with extremely high accuracy [70–95%] by considering the brain mosaic as a whole', even though there is variation within the categories men and women (which they analogize to 'breeds of dogs') and no 'singular physical characteristic reliably distinguishes cats from dogs' or men's from women's brains (Chekroud et al., 2016). Where the univariate approach failed to show a distinction between men and women (or cats and dogs), ML rescues the theory that they are distinct by finding multivariate patterns that are predictive of sex.

The dogs vs cats and species of monkeys analogies that these authors make, like the gorillas, children and 'savages' analogies Le Bon made 137 years earlier, offer insight into how the SIM conceives of sex difference: men and women are analogous to different species. Their theory of sex difference is not what we may be familiar with from introductory statistics courses, different averages from largely overlapping populations. Instead, they envision a categorical distinction in kind. ML classifiers promise to mathematically separate such kinds even when they seem like indistinct, overlapping mosaics.

Most sexual dimorphism papers are not direct rebuttals of specific feminist publications, but they do set out to show that human brains come in two categorically distinct forms, with titles like 'Combined structural and resting-state functional MRI analysis of sexual dimorphism in the young adult human brain' (Wang et al., 2012). Seventeen (68%) of the sexual dimorphism papers include statements in their introduction asserting that sexual dimorphism in the human brain is already an established fact. In these, sex differences are both the explicit starting premise and the expected finding. Baldinger-Melich et al. (2020) make this their very first sentence: 'Sex differentiation of all somatic tissues, including the brain, is driven both by direct genetic influences and gonadal hormones with only minimal environmental effects'. Others assert that 'The existence of organic differences in the brain, attributable to sex, has been well-replicated and reviewed at an aggregate level' (Anderson et al., 2019, p. 1496) and that 'Previous studies … have revealed consistent differences in whole brain tissue volume between the sexes' (Sepehrband et al., 2018, p. 217). By finding that brains can be classified as male and female, ML does not challenge or advance what the authors claim is established knowledge about sex and brains. Instead, it reestablishes that knowledge, and the SIM's need to reestablish that knowledge betrays its unsettled status.

ML methods were introduced into neuroimaging in the early 2000s explicitly as a solution to the growing failure of the dominant approach (Haxby, 2012). The first paper to use them for showing sex difference said so clearly in its abstract: using ML 'classification methods, [we] can detect subtle and spatially complex patterns of morphological group differences which are often not detectable by…[univariate] methods [that] analyze morphological measurements voxel-by-voxel and do not consider the entirety of the data simultaneously' (Lao et al., 2004, p. 46). Twenty out of 26 papers published before 2019 have statements like this one, saying ML is promising because it will find sex differences where other methods have failed. By 2019, the use of ML in neuroimaging was more established and fewer papers made explicit comparisons to univariate approaches.

Some use diagrams with made-up data to explain this point (e.g. Huf et al., 2014; Lao et al., 2004; Wachinger et al., 2015). One example is Figure 2, reprinted from

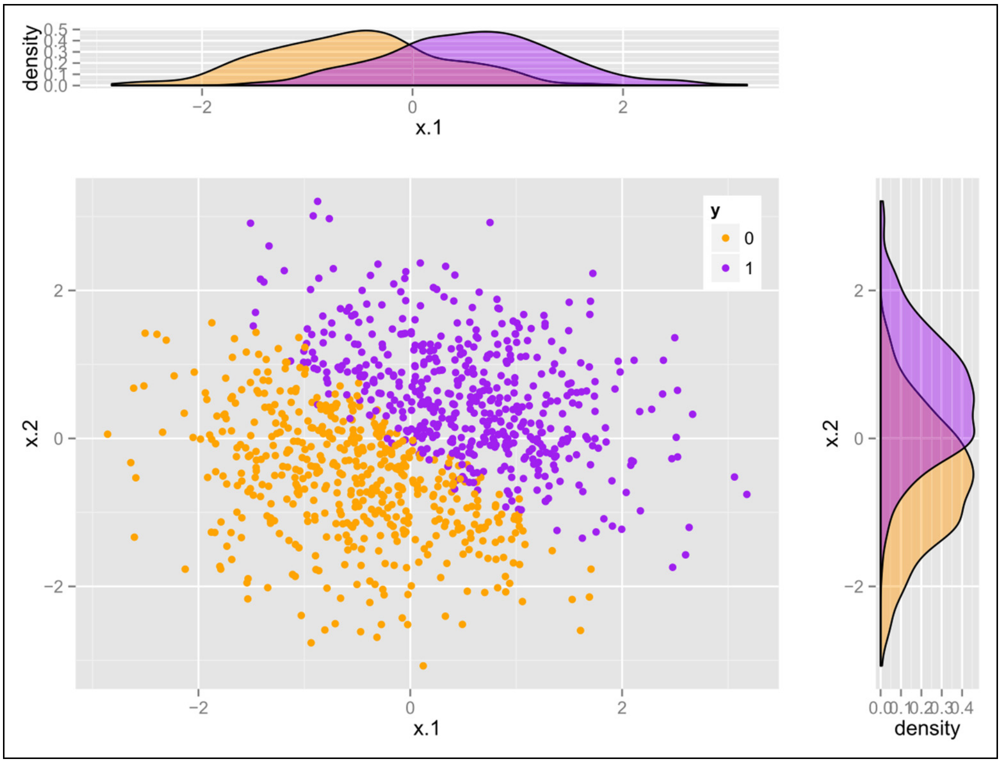

Figure 2. Rosenblatt's (2016) hypothetical data justifying the use of machine learning.

Rosenblatt's (2016) response to Joel in *PNAS*. Rosenblatt argues that even if no measures of brains appear categorically distinct in univariate analysis, men's and women's brains might still be categorically distinct in multivariate analysis. He illustrates the point with made-up data showing an almost perfect separation between two groups that seem to overlap if we look at only one variable at a time. This is the promise of ML for authors seeking to prove sexual dimorphism. Despite using over 100 variables, Rosenblatt's ML model is unable to produce the clean categorical distinction promised by his hypothetical diagram with two variables: it yields only 80% accuracy, indicating substantial overlap between 'male' and 'female' brains. Nevertheless, in a perfect example of McQuillan's (2018) 'machinic neoplatonism', Rosenblatt concludes 'given our empirical evidence and the *multivariate intuition depicted above*, we cannot help but disagree with [Joel]…Brains are indeed typically male or typically female' (Rosenblatt, 2016, p. E1966 emphasis added). The theoretical appeal of the algorithm, depicted in a figure with made-up data, is compelling enough to override the actual results. ML classification brings with it an intuition that lets scientists 'see' a categorical difference *in spite of* the data and results they observe. That intuition is helpful to SIM actors arguing for categorical differences between human groups.

This multivariate intuition of ML classifiers is radically different from the univariate, frequentist statistical approaches that dominated neuroscience before. Traditional approaches are about group properties that exist only in aggregate. Whether scientists are working with regression coefficients, two-sample *t*-tests, or ANOVAs, there is a general understanding that the groups being compared are not categorically distinct, with their properties true of all members, but rather overlapping distributions whose properties are averages of diverse members. As a result, neuroscientists generally use 'risk-thinking', focusing, for example, 'on who *could be* violent rather than who *was* or *is* violent' (Rollins, 2021, p. 15 emphasis original). This puts SIM members in a tricky place. In order to make their point that sex is *categorical* in the strong sense that they mean using traditional statistics, they need to elide the messy overlap between groups and misrepresent the statistical 'average man' as universal (Igo, 2007; Lockhart, 2020). Such analyses emphasized significance testing for the existence of differences over the effect sizes of differences, which are typically very small (Eliot et al., 2021).

ML offers a way to escape this bind and speak confidently about the categorical difference between groups. ML *is* about individual predictions and labels. Its focus is on 'y-hats' (predicting facts about individuals like their sex or whether they commit violent acts) rather than 'beta-hats' (risk factors, group averages) (Salganik, 2017). Brennan, Wu, and Fan lay out the logic: 'If the [ML model] can successfully classify biological sex significantly above chance using these features [brain scan data], this would demonstrate that dimorphism exists in the brain' (Brennan et al., 2021, p. 2). This is fundamentally different from the frequentist approach, as Anderson et al. explain: 'while on-average differences in these regions have been recognized previously… it is conceptually important that the current work has demonstrated multivariate [ML] models that approach something closer to *truly dimorphic* patterns, effectively differentiating individuals into categorical groups based on intrinsic [brain] structural networks' (Anderson et al., 2019, p. 1503 emphasis added). While the averages of the frequentist approach lend themselves to questions of magnitude of difference, classification accuracies focus only on the existence of difference.

To see how ML better fits theories of categorical difference, consider these descriptions of the most popular ML algorithm in the brain sex literature, the SVM:

> [SVM] is a classification algorithm that attempts to find a hyperplane which best separates binary classes (male and female) in hyperspace of predictive features (Brennan et al., 2021, p. 2).

> the algorithm finds a hyperplane (i.e. a high dimensional plane) that maximizes the margin between the training samples of both classes.… Kernel functions can be applied if the data are not linearly separable in the original space and allows for group classification based on non-linear effects (Anderson et al., 2019, p. 1499)

> A hyperplane (or hypersurface) is determined that optimally separates the two groups of samples (Lao et al., 2004, p. 49).

> linear SVM… assigns a decision value to each subject, reflecting the distance between a given test subject's [brain] images and the hyperplane separating the two groups (Feis et al., 2013, p. 254).

These are fairly standard descriptions of SVMs, echoing what one hears in ML classes and the primary SVM literature. They all imagine something like the image in Figure 2 and say that SVM will draw a line ('hyperplane' is the name for a line in data with many variables instead of just two) across the diagonal separating the data points representing brains from men and women. Two of the quoted statements flatly assert that such a line is found, implying both that the task is possible and that the algorithm will succeed. One says SVM 'attempts' to find such a line but does not elaborate on what failure would look like. Another suggests 'kernel functions' can be used to make SVM find a dividing line even when no straight line could separate the data points for men and women.

A method that can do what SVM promises has substantial allure for research on categories. If we can really draw a line separating two groups, as implied by these descriptions of SVM and the illustrations of made-up data, then these scientists claim it would be strong evidence of those groups' binary, categorical nature. Fisher thought so in 1936 when he published a paper on Linear Discriminant Analysis in the *Annals of Eugenics*. Like SVM, LDA aims to find lines separating groups, and Fisher noted its use in craniometry and for sex differences (Fisher, 1936, p. 179). While SVM is by far the most popular method for predicting sex from brains, used in 33 out of 53 papers, I also found four papers between 2013 and 2020 using LDA to classify brains by sex, often alongside SVM, deep neural networks, and other ML models, demonstrating continuity of reasoning about sex categories over time that scientists engaged in contemporary SIM struggles profess they are unaware of (Lockhart, 2020).

Yet in practice, these separating lines are fiction. SVM generally does not converge on real data without a C parameter that allows some data points from each group to be on the 'wrong side' of the dividing line. The simplest evidence of this is the results in the papers themselves. None of the 53 papers I collected presents 100% accuracy in predicting whether a brain is male or female. There are always some brains on the 'wrong side' of the dividing line. The top reported accuracies from each paper range from 51% to 98%, averaging 82% and showing little improvement over time or as sample sizes increase. Indeed, others have argued that the lack of relationship between sample size and number of sex differences 'discovered' is evidence of reporting bias in this literature (David et al., 2018). Widespread design flaws in this research have also led to inflated accuracies, including failure to properly control for head size and other confounds (Dhamala et al., 2020; Eliot et al., 2021; Linn et al., 2016; Sanchis-Segura et al., 2020), failure to test on large enough and independent samples (Glocker et al., 2019; Huf et al., 2014; Varoquaux, 2018; Zoubi et al., 2020), and failure to think through model assumptions (Figure 3) (Bzdok, 2017; Carlson et al., 2018; Schulz et al., 2020).

But even taking the reported accuracies at face value[2], how are accuracy scores mostly in the 1970s and 1980s presented as success in papers premised on the categorical separability of male and female brains? If a model is 82% accurate, then it puts 18% of people into the 'wrong' sex category. It fails to separate male from female. As is always the case with scientific facts, constructing the fact of sexual dimorphism takes more work after the results. Importantly, sex difference scholars leverage the specific categorical design of ML to rescue their theory of categorical sex.

In a typical turn of phrase, Nieuwenhuis et al. tell us that 'sex was predicted with significant accuracy (89%; $p < 0.001$)' (Nieuwenhuis et al., 2017:, p. 246). In ML, one does not typically compute statistical significance. Yet nearly every neuroscience paper using ML reports that their model achieved significant accuracy. This bolsters an argument that the ML model is correct, and since the model embodies the SIM's ontology of sex as two distinct, separable categories, if the model is correct, then so is their theory. For example, Chekroud et al. report 95% accuracy before controlling for confounding variables, a number that stands on its own to show the model is correct. But after controls reduce accuracy to 70%, the authors bring in significance testing, arguing that the models 'remained significant' and thus still represent the correct understanding of sex in brains (Chekroud et al., 2016, p. E1968). Van Putten, Olbrich, and Arns do something similar, initially setting up a high bar and then using significance testing to lower it: 'We have excellent skills to extract sex from visual assessment of human faces, but assessing sex from human brain rhythms seems impossible', so they calculated 'a significance threshold for the classification accuracy of 63%', meaning any accuracy above 63% is significant (Van Putten et al., 2018, p. 1). Through

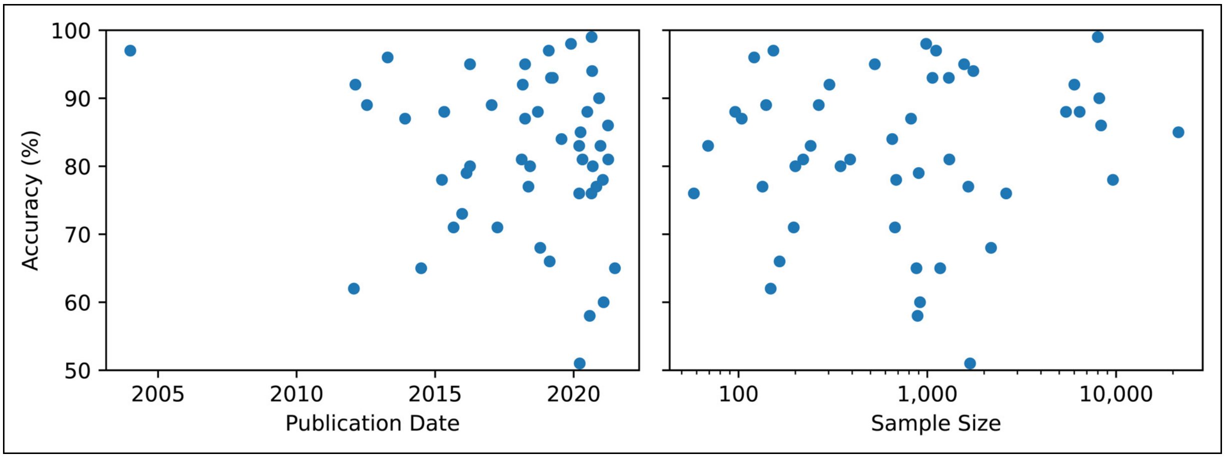

Figure 3. Reported accuracy of sex prediction from brain scans by publication date and sample size.

significance testing, scientists can assert that their categorical models of sex are correct despite the inaccuracies, because the models are ‘significant’, which is often (mis) understood as shorthand for ‘true’ or ‘correct’.

Each ML method used in this literature has its own narrative like SVM, and they share many of the same promises and challenges. Since 2018, neural networks have become increasingly popular in this area, used in 14 papers (26%). Deep learning adds further credibility to brain sex research by promising to bypass the theoretical assumptions necessary for other types of models and operate directly on the ‘raw’ brain image data to ‘learn the kernel’ or functional form of the relationship between brain data and sex (Bzdok, 2017; Carlson et al., 2018; Schulz et al., 2020, p. 2). Such rhetoric is a form of ‘enchanted determinism’, in which ML exceeds and replaces human capacity to understand patterns, further adding to its credibility and preventing critique by SIM opponents (Campolo and Crawford, 2020). Yet ‘raw data is an oxymoron’, especially in neuroimaging data, which is always heavily processed (normalized, aligned, warped, motion-corrected, skullstripped, etc.) (others have described this extensively, e.g. Jackson et al., 2013; Roskies, 2008).

Symbiotic scientists using sex as proof of concept

The second largest group of papers (n = 16, or 30%) are what I call proof-of-concept papers. These papers are interested in demonstrating a new method for analyzing neuroimaging data, and they use sex classification as their example use case. Unlike sexual dimorphism papers, which are published in a mix of neuroscience and general-interest journals, proof-of-concept papers are divided between neuroscience journals and computer science conferences.

Proof-of-concept papers say almost nothing about sex/ gender. For example, the paper ‘On the generalizability of resting-state fMRI machine learning classifiers’ (Huf et al., 2014) is fundamentally disinterested in sex. The authors mention sex only once, parenthetically, saying ‘classification accuracies of up to 0.8 (using sex as the target variable) could be achieved’ with their proposed method (Huf et al., 2014, p. 1). Some proof-of-concept papers are explicit about why they use sex to demonstrate their method: ‘we use the subjects’ gender label [male/ female] as the predicting label because gender is the golden standard in the neuroimaging field and does not include subjective factors’ (Yuan et al., 2018, p. 49926). Others concur: ‘the classification problem is relatively easy, as sex can be unequivocally determined and brain sexual dimorphisms is [sic] well established’ (Nieuwenhuis et al., 2017, p. 248). Whether writing in computer science or neuroscience, they assume their audience does not need to be convinced that brain sexual dimorphism is an established scientific fact. Rather than framing the performance of their ML models as evidence for the scientific fact of brain sex dimorphism, the fact is taken as given and the performance of the models is interpreted as evidence of the models’ quality. Indeed, these authors use sex difference as a competition: each paper compares its performance classifying sex to prior papers, showing the new method better separates men and women’s brains. For them, sex difference is not a theory about brains to be tested or proven, but rather something to be engineered and maximized.

The sex difference SIM and this symbiotic group of computer science authors are part of the same publishing and citation ecosystem. Proof-of-concept papers are published and cited contemporaneously with sexual dimorphism papers in overlapping journals. It is not that scientists first demonstrated that human brains come in two discrete categories (male and female) and then used these validated categories to test new methods. Instead, in a circular chain of logic, some scientists are using the accuracy of ML to debate whether human brains are sexually dimorphic at the same time that others are using sex dimorphism as a gold standard to test whether algorithms are accurate. Proof-of-concept papers, seemingly unaware of the controversy around brain sex dimorphism, nevertheless contribute to the accretion of evidence for it by adding to the list of publications ‘finding’ the dichotomy. By not acknowledging the controversy, such papers do more to shore up the sex binary than even the dimorphism papers, as the latter usually mention mixed results in their literature reviews and research limitations.

Proof-of-concept papers are also generally aimed at different audiences. Sexual dimorphism papers are often written for general science audiences to convince them of sex differences, while proof-of-concept papers are often written for computer scientists to share new methods for classification. Even among papers in the same journal, such as *NeuroImage*, the introductions, literature reviews, conclusions, and significance statements make it clear they have different audiences. Dimorphism papers are written for readers interested in the nature of sex in brains, while proof-of-concept papers are written for readers interested in the methods of analysis. This later audience is less likely to be aware that brain sex classification is controversial among domain experts and thus more likely to take claims about it at face value. Thus, ML’s community of methodologists also serves as a resource for spreading the sex difference SIM’s core idea.

Countermovement adaptation to ML

The sex difference SIM faces opposition from a feminist countermovement, which has also begun to adopt ML. They design research methods using ML that materialize their understandings of sex as nonbinary, contingent on researcher choice, or not a primary organizing principle of brains. I found four examples of this, none of which have so far been taken up or become institutions of neuroscientific practice.

In the simplest approach, authors exploit the interpretive flexibility of ML, arguing that the same data, methods, and results actually encode continuous, nonbinary understandings of sex rather than discrete, binary categories. Zhang et al. (2021) argue that 'The Human Brain is Best Described as Being on a Female/Male Continuum' by taking advantage of the fact that SVM outputs predicted probabilities. For example, SVM might output 74% probability that a brain image came from a man based on its distance from the male/female dividing line. Typically, researchers apply a threshold and say anything over, for example, 50% probability is male and anything below that is female, thus turning a continuous measure into a categorical one and separating male and female. Zhang et al. (2015) skip this step. Building on Joel's mosaic brain theory, they assume all brains are a mix of 'male' and 'female features' and then use the probabilities from their SVM to place brains on a spectrum rather than in binary categories. They show that the predicted probability of being male in brains is correlated with self-reports of gendered behavior, beliefs, and psychological metrics. Zhang et al. conclude that 'The moderate classification accuracy [78%] of the multivariate classifier indicated that the brain functional architecture was unlikely to be conceptualized as binary, as is the case with biological sex, but was more likely to be continuously represented on a brain gender spectrum' (Zhang et al., 2021, p. 11). Further highlighting their SIM agenda, they argue that 'androgyny' in brains and behavior 'is advantageous for mental health' (Zhang et al., 2021, p. 11). In other words, the sex binary is not just scientifically inaccurate but actively harmful to people.[3]

Sanchis-Segura and colleagues (2020, 2022) take a different approach. They still use the same kind of data (structural MRI) and ML classification methods (SVM, neural networks, and 10 others). But rather than building and trusting a single model, or building multiple models and reporting the results from the best-performing one, they construct a variety of models with different measurement assumptions, then compare the results to understand how assumptions about sex influence the findings of neuroscience. They demonstrate that whether and how researchers correct for head size (total intracranial volume, TIV) and how they operationalize sex categories both have a substantial impact on the accuracy of ML models. One way of controlling for TIV reduces the average accuracy of models to 57%[4], and allowing an indeterminate sex category reduces accuracy to 43%.

Choices like how to handle head size depend on researchers' assumptions. Sanchis-Segura et al. (2020, p. 12953) point out that others have observed large discrepancies between accuracies using corrected and uncorrected data but chose to emphasize the uncorrected results. So what one believes about the importance of head size (or, cynically, what results a SIM member wants to show) is a critical aspect in the choice of methods. Studies feeding 'raw' brain scan data into ML models are not as atheoretical as they appear; they are implicitly taking the contested stance that raw size matters. By comparing models, Sanchis-Segura's team is able to avoid Rosenblatt's (2016) trap that classification models can only support claims of categorical difference. Their comparative use of ML demonstrates the role of researcher assumptions in findings of difference, helping to de-naturalize the sense of objectivity and undo the enchanted determinism around findings of sex difference.

In the most methodologically innovative paper, Joel et al. (2018) look beyond classification methods to other tools from ML: anomaly detection and clustering. Anomaly detection is an intuitive approach to the categorical difference question because it tests whether 'brain type(s) typical of females [are] also typical of males' (Joel et al., 2018, p. 5). These models are given training data for something (e.g. brain scans from women) and then asked to label whether new data (e.g. brain scans from men) appear typical of the data they have already seen, or anomalous. Going a step further, Joel's team takes up their critics' monkey species analogy (Del Giudice et al., 2015) and shows that anomaly detection can differentiate the faces of monkey species but not brains of human men and women. Sex difference, they conclude, is not like species difference.

Joel et al. (2018) also use clustering algorithms (which include tools like topic modeling). The logic here is straightforward: if brains really come in distinct forms according to sex categories, then tools designed to cluster the data into categorical groups should return groups that are largely sex-segregated. Unlike classification methods, where the computer is tasked with 'discriminating male and female brains', clustering methods are tasked with finding the best division of brain data into groups without considering sex. When Joel's team compares the groups produced by the model with participant sex, they find the poor correspondence. By using ML in which categorical sex difference is not baked into the design as an outcome variable, Joel's team is able to imagine and find alternate ways of understanding sex and variation in brains.

Conclusions

At a time when neuroscientists' confidence in finding constructs such as sex, intelligence, and criminality in brains was wavering, ML's promise to find them using relationships among brain variables that neuroscientists had previously considered independently offered an enticing solution. Once multivariate ML approaches were adopted, even sex difference proponents admitted that univariate approaches failed to find sex in brains. No individual brain features distinguish men and women, and no brain is composed of only 'male' or 'female characteristics' (Chekroud et al., 2016). ML serves as a resource by enabling the continued publication of papers advancing the SIM's position after the collapse of the prior methodological approach.

ML also offered neuroscientists a fundamentally different way of thinking about groups of people. Rather than their longstanding statistical conception of overlapping distributions and risk factors, neuroscientists began imagining groups of people as clouds of data points that could be cleanly separated into distinct categories by a line. Neuroscientists in the SIM for sex differences saw an opportunity: this new conception of group difference was much closer to their core idea that brains come in two distinct forms, male and female, dog and cat. ML classification technology institutionalizes the sex difference SIM's main idea, shifting how scientists in their field frame group difference to more favorable terms. SIM members explicitly comment and capitalize on this in their publications, saying it offers better evidence for their idea of 'truly categorical difference' than the previous quantitative paradigm (Anderson et al., 2019). In this way, the research method also serves as a resource for the SIM by shifting the field's framing of group difference to align with the SIM's framing. While prior work focuses on SIMs' use of framing to make their motivations seem important to their peers (Frickel and Gross, 2005), I show that a research method can be used to frame a SIM's substantive idea more palatable by institutionalizing it.

The anti-essentialist countermovement has attempted to use ML for the same end, institutionalizing their view of brain sex through strategic use and interpretation of ML in their research methods. With only four papers using ML, each in a different way, the countermovement has materialized its values in the technology (Bucher, 2016), but it has not been successful in institutionalizing those ideas because as one-off technologies, they do not yet pattern the activities of scientific reproduction (Pinch, 2008).

The use of ML by partisans of the sex difference SIM also brought in and allied a new constituency, computer scientists, who took up the SIM's idea (sex in brains is binary and easily detectable) for their own ends. The two groups developed a symbiotic relationship, propelling both forward. Brain sex difference advocates offered computer scientists a new task to optimize and compete over. The computer scientists, in turn, gave back a series of publications engineering ever more certain differences between men's and women's brains.

It remains to be seen how widely ML will be used to institutionalize strict categorical understandings of human groups in other scientific fields such as genetics, psychology, or even sociology, as well as how widely it will be taken up by the SIMs arguing about human difference in race, intelligence, sexuality, and disability. Attempts have been made on all of these fronts and more—including classifying political party and criminality—but most cases have been fairly isolated (Arcas et al., 2017; Lockhart and Jacobs, 2021). Even within neuroscience of sex, the existing interpretive flexibility around ML suggests we may see future changes in which conceptions of difference ML institutionalizes and the extent of their spread.

Nevertheless, the specific technologies of ML encode new quantitative reasoning(s) about group difference which will undoubtedly propel some SIMs and demobilize others (Krippner and Hirschman, 2022). If we wish to understand these movements or broader scientific claims about important social categories, it is critical that we examine not only how research methods are deployed and interpreted, but also how they work as technologies and how they embed and obscure contested ideas and values (Bucher, 2016; Pinch, 2008). Such analyses of methods extend the more developed sociological literature examining how measurement embeds ideas about things like gender and shapes what is knowable (Cicourel, 1964; Westbrook and Saperstein, 2015).

Acknowledgments

Chalem Bolton, Elizabeth Bruch, Erin Cech, Marina DiMarco, Parker Koch, Jason Owen-Smith, Joy Rohde, Mike Thompson-Brusstar, Donna, and members of the GenderSci Lab provided helpful feedback on earlier versions of this manuscript. This research was supported in part by an NICHD training grant to the Population Studies Center at the University of Michigan (T32HD007339).

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grant number T32HD007339).

ORCID iD

Jeffrey W. Lockhart https://orcid.org/0000-0003-1519-9588

Notes

1. The vast majority of natural science papers are written in English, even in countries where authors' and audiences' first language is generally not English (Di Bitetti and Ferreras, 2017). The authors of the papers in my sample are based in 22 countries on five continents.
2. At least one of the papers retracted its findings after discovering a coding error (Ecker, 2019).
3. Such spectrum-based accounts of brain sex, however, have recently been used to argue in favor of essentialist sex

differences. One paper published after my data collection argues that a significant correlation between sex score and intelligence shows that differences in intelligence are based on sex, such that men with low intelligence have brains that the algorithm places on the female end of the spectrum, and high-intelligence women have brains on the male end of the spectrum (Kim et al., 2022).

4. There has long been debate about the best way to adjust for TIV. This is a methodological argument about what approach should be standard, but it is also a SIM struggle to institutionalize a method aligned with their framing of sex as a factor that either does or does not remain important independent of things like head size.

References

Anderson ML (2014) *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press.

Anderson NE, Harenski KA, Harenski CL, et al. (2019) Machine learning of brain gray matter differentiates sex in a large forensic sample. *Human Brain Mapping* 40(5): 1496–1506.

Arcas BA, Mitchell M and Totorov A (2017) Physiognomy's new clothes. In: *Medium*. Available at: https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a (accessed 28 September 2020).

Arribas-Ayllon M, Bartlett A and Featherstone K (2010) Complexity and accountability: The witches' brew of psychiatric genetics. *Social Studies of Science* 40(4): 499–524.

Baldinger-Melich P, Urquijo Castro MF, Seiger R, et al. (2020) Sex matters: A multivariate pattern analysis of sex- and gender-related neuroanatomical differences in cis- and transgender individuals using structural magnetic resonance imaging. *Cerebral Cortex* 30(3): 1345–1356.

Benjamin R (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*, 1st edition Medford, MA: Polity.

Brennan D, Wu T and Fan J (2021) Morphometrical brain markers of sex difference. *Cerebral Cortex* 31(8): 3641–3649.

Bucher T (2016) Neither black nor box: Ways of knowing algorithms. In: Kubitschko S and Kaun A (eds) *Innovative Methods in Media and Communication Research*. Cham: Springer International Publishing, 81–98.

Bzdok D (2017) Classical statistics and statistical learning in imaging neuroscience. *Frontiers in Neuroscience* 11: 1–23.

Cahill L (2014) Equal ≠ the same: sex differences in the human brain. *Cerebrum: The Dana Forum on Brain Science* 2014: 5.

Campolo A and Crawford K (2020) Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society* 6: 1–19.

Carlson T, Goddard E, Kaplan DM, et al. (2018) Ghosts in machine learning for cognitive neuroscience: Moving from data to theory. *NeuroImage* 180: 88–100.

Chekroud AM, Ward EJ, Rosenberg MD, et al. (2016) Patterns in the human brain mosaic discriminate males from females. *Proceedings of the National Academy of Sciences* 113(14): E1968–E1968.

Chun WHK (2021) *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge, MA, USA: MIT Press.

Cicourel AV (1964) *Method and Measurement in Sociology*. New York, NY: Free Press.

Clayton A (2020) How eugenics shaped statistics. *Nautilus*, 28 October. Available at: http://nautil.us/issue/92/frontiers/how-eugenics-shaped-statistics (accessed 29 October 2020).

Collins H (1992) *Changing Order: Replication and Induction in Scientific Practice*. Chicago, IL: University of Chicago Press. Available at: https://press.uchicago.edu/ucp/books/book/chicago/C/bo3623576.html (accessed 21 May 2022).

David SP, Naudet F, Laude J, et al. (2018) Potential reporting bias in neuroimaging studies of sex differences. *Scientific Reports* 8(1): 6082.

Del Giudice M, Lippa R, Puts D, et al. (2015) Mosaic brains? A methodological critique of Joel et al. Behavioral and Psychological Sex Differences. doi: 10.13140/RG.2.1.1038.8566.

Dhamala E, Jamison KW, Sabuncu MR, et al. (2020) Sex classification using long-range temporal dependence of resting-state functional MRI time series. *Human Brain Mapping* 41(13): 3567–3579.

Di Bitetti MS and Ferreras JA (2017) Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications. *Ambio* 46(1): 121–127.

Ecker C (2019) Notice of retraction and replacement: Ecker et al. Association between the probability of autism spectrum disorder and normative sex-related phenotypic diversity in brain structure. *JAMA Psychiatry* 74(4): 329–338.

Eliot L, Ahmed A, Khan H, et al. (2021) Dump the "dimorphism": Comprehensive synthesis of human brain studies reveals few male-female differences beyond size. *Neuroscience & Biobehavioral Reviews* 125: 667–697.

Emerson RM, Fretz RI and Shaw LL (2011) *Writing Ethnographic Fieldnotes*. Chicago, IL: University of Chicago Press.

Epstein S (2007) *Inclusion: The Politics of Difference in Medical Research*. Chicago: University of Chicago Press.

Eubanks V (2017) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, 1st edition New York, NY: St. Martin's Press.

Fausto-Sterling A (2005) The bare bones of sex: Part 1-sex and gender. *Signs* 30(2): 1491–1527.

Feis D-L, Brodersen KH, von Cramon DY, et al. (2013) Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. *NeuroImage* 70: 250–257.

Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(2): 179–188.

Fitsch H (2021) Reflections on binary sex/gender categorization in magnetic resonance tomography and its future challenges. *Frontiers in Sociology* 6: 1–6.

Fourcade M and Healy K (2016) Seeing like a market. *Socio-Economic Review* 15(1): 9–29.

Fourcade M and Healy K (2017) Categories all the way down. *Historical Social Research* 42(1): 286–296.

Frickel S and Gross N (2005) A general theory of scientific/intellectual movements. *American Sociological Review* 70(2): 204–232.

Glocker B, Robinson R, Castro DC, et al. (2019) Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv:*1910.04597 *[cs, eess, q-bio]*. Available at: http://arxiv.org/abs/1910.04597 (accessed 17 April 2021).

Gould SJ (1996) *The Mismeasure of Man. Rev. and Expanded*. New York: Norton.

Haxby JV (2012) Multivariate pattern analysis of fMRI: The early beginnings. *Neuroimage* 62(2): 852–855.

Huf W, Kalcher K, Boubela RN, et al. (2014) On the generalizability of resting-state fMRI machine learning classifiers. *Frontiers in Human Neuroscience* 8: 1–11.

Igo SE (2007) *The Averaged American: Surveys, Citizens, and the Making of a Mass Public*. Cambridge, MA: Harvard University Press.

Jackson V, Rosenberg D, Williams TD, et al. (2013) *'Raw Data' Is an Oxymoron (ed. L Gitelman)*. Cambridge, MA: The MIT Press.

Joel D, Berman Z, Tavor I, et al. (2015) Sex beyond the genitalia: The human brain mosaic. *Proceedings of the National Academy of Sciences* 112(50): 15468–15473.

Joel D, Persico A, Salhov M, et al. (2018) Analysis of human brain structure reveals that the brain "types" typical of males are also typical of females, and vice versa. *Frontiers in Human Neuroscience* 12: 1–18.

Keyes O (2018) The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW): 1–22.

Keyes O, Hitzig Z and Blell M (2021) Truth from the machine: Artificial intelligence and the materialization of identity. *Interdisciplinary Science Reviews* 46(1–2): 158–175.

Kim K, Joo YY, Ahn G, et al. (2022) The sexual brain, genes, and cognition: A machine-predicted brain sex score explains individual differences in cognitive intelligence and genetic influence in young children. *Human Brain Mapping* 43(12): 3857–3872.

Krippner GR and Hirschman D (2022) The person of the category: the pricing of risk and the politics of classification in insurance and credit. *Theory and Society* 51: 685–727.

Lao Z, Shen D, Xue Z, et al. (2004) Morphological classification of brains via high-dimensional shape transformations and machine learning methods. *NeuroImage* 21(1): 46–57.

Laqueur TW (1990) *Making Sex: Body and Gender from the Greeks to Freud*. Cambridge, Mass: Harvard University Press.

Latour B (2005) *Reassembling the Social*. New York: Oxford University Press.

Linn KA, Gaonkar B, Doshi J, et al. (2016) Multivariate pattern analysis and confounding in neuroimaging. *The International Journal of Biostatistics* 12(1): 31–44.

Lockhart JW (2020) 'A large and long standing body': Historical authority in the science of sex. In: Valencia-García LD (ed) *Far Right Revisionism and the End of History: Alt/Histories*. New York: Routledge, 359–386.

Lockhart JW (2021) Paradigms of sex research and women in STEM. *Gender & Society* 35(3): 449–475.

Lockhart JW and Jacobs AZ (2021) Scientific Argument with Supervised Learning. NeurIPS 2021 AI for Science Workshop. 13 December, 2022 1–5.

McQuillan D (2018) Data science as machinic neoplatonism. *Philosophy & Technology* 31(2): 253–272.

Morning A (2014) And you thought we had moved beyond all that: Biological race returns to the social sciences. *Ethnic and Racial Studies* 37(10): 1676–1685.

Nieuwenhuis M, Schnack HG, van Haren NE, et al. (2017) Multi-center MRI prediction models: Predicting sex and illness course in first episode psychosis patients. *NeuroImage* 145: 246–253.

Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY: NYU Press.

Pinch T (2008) Technology and institutions: Living in a material world. *Theory and Society* 37(5): 461–483.

Rippon G (2019) *The Gendered Brain: The New Neuroscience That Shatters the Myth of the Female Brain*. London: The Bodley Head.

Rollins O (2021) *Conviction: The Making and Unmaking of the Violent Brain*. Stanford: Stanford University Press.

Rosenblatt JD (2016) Multivariate revisit to "sex beyond the genitalia". *Proceedings of the National Academy of Sciences* 113(14): E1966–E1967.

Roskies AL (2008) Neuroimaging and inferential distance. *Neuroethics* 1(1): 19–30.

Salganik M (2017) *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Sanchis-Segura C, Aguirre N, Cruz-Gómez ÁJ, et al. (2022) Beyond "sex prediction": Estimating and interpreting multivariate sex differences and similarities in the brain. *NeuroImage* 257: 119343.

Sanchis-Segura C, Ibañez-Gual MV, Aguirre N, et al. (2020) Effects of different intracranial volume correction methods on univariate sex differences in grey matter volume and multivariate sex prediction. *Scientific Reports* 10(1): 1–15.

Sanz V (2017) No way out of the binary: A critical history of the scientific production of sex. *Signs* 43(1): 1–27.

Scheuerman MK, Pape M and Hanna A (2021) Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data & Society* 8(2): 20539517211053710.

Schilt K and Westbrook L (2009) Doing gender, doing heteronormativity: "gender normals," transgender people, and the social maintenance of heterosexuality. *Gender & Society* 23(4): 440–464.

Schulz M-A, Yeo BTT, Vogelstein JT, et al. (2020) Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications* 11(1): 1.

Seaver N (2018) What should an anthropology of algorithms do? *Cultural Anthropology* 33(3): 375–385.

Sepehrband F, Lynch KM, Cabeen RP, et al. (2018) Neuroanatomical morphometric characterization of sex differences in youth using statistical learning. *NeuroImage* 172: 217–227.

Sudai M, Borsa A, Ichikawa K, et al. (2022) Law, policy, biology, and sex: Critical issues for researchers. *Science (New York, N.Y.)* 376(6595): 802–804.

van Putten MJAM, Olbrich S and Arns M (2018) Predicting sex from brain rhythms with deep learning. *Scientific Reports* 8(1): 3069.

Varoquaux G (2018) Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* 180: 68–77.

Wachinger C, Golland P, Kremen W, et al. (2015) Brainprint: A discriminative characterization of brain morphology. *NeuroImage* 109: 232–248.

Wade L (2013) The new science of sex difference. *Sociology Compass* 7(4): 278–293.

Wang L, Shen H, Tang F, et al. (2012) Combined structural and resting-state functional MRI analysis of sexual dimorphism in the young adult human brain: An MVPA approach. *NeuroImage* 61(4): 931–940.

Westbrook L and Saperstein A (2015) New categories are not enough: Rethinking the measurement of sex and gender in social surveys. *Gender & Society* 29(4): 534–560.

Yuan L, Wei X, Shen H, et al. (2018) Multi-Center brain imaging classification using a novel 3D CNN approach. *IEEE Access* 6: 49925–49934.

Zhang Y, Luo Q, Huang C-C, et al. (2021) The human brain is best described as being on a female/male continuum: Evidence from a neuroimaging connectivity study. *Cerebral Cortex* 31(6): 3021–3033.

Zoubi OA, Misaki M, Tsuchiyagaito A, et al. (2020) Predicting Sex from Resting-State fMRI Across Multiple Independent Acquired Datasets. *BioRxiv*. doi: 10.1101/2020.08.20.259945.