Oxford Internet Institute, University of Oxford

Assignment Cover Sheet

Assignment	Fairness, Accountability and Transparency in Machine Learning: Summative Essay
Term	Hilary Term 2024
Title/Question	Auditing Ancestry: Considering the Algorithmic Ecology of the AncestryDNA Ethnicity Estimate
Word Count	4846

By placing a tick in this box [✓] I hereby certify as follows:

(a) This thesis or coursework is entirely my own work, except where acknowledgments of other sources are given. I also confirm that this coursework has not been submitted, wholly or substantially, to another examination at this or any other University or educational institution;

(b) I have read and understood the Education Committee's information and guidance on academic good practice and plagiarism at `https://www.ox.ac.uk/students/academic/guidance/skills?wssl=1`.

(c) I agree that my work may be checked for plagiarism using Turnitin software and have read the Notice to Candidates which can be seen at: `http://www.admin.ox.ac.uk/proctors/turnitin2w.shtml`, and that I agree to my work being screened and used as explained in that Notice;

(d) I have clearly indicated (with appropriate references) the presence of all material I have paraphrased, quoted or used from other sources, including any diagrams, charts, tables or graphs.

(e) I have acknowledged appropriately any assistance I have received in addition to that provided by my [tutor/supervisor/adviser].

(f) I have not sought assistance from a professional agency;

(g) I understand that any false claims for this work will be reported to the Proctors and may be penalized in accordance with the University regulations.

Please remember:

- To attach a second relevant cover sheet if you have a disability such as dyslexia or dyspraxia. These are available from the Higher Degrees Office, but the Disability Advisory Service will be able to guide you.

Auditing Ancestry: Considering the Algorithmic Ecology of the AncestryDNA Ethnicity Estimate

1 Introduction

1.1 AncestryDNA's Ethnicity Estimate

Whereas many direct to consumer (DTC) genetic testing services evolved from commercial healthcare entities, Ancestry was founded in 1983 as a publishing company. The objective was to facilitate genealogical research at an accessible level by providing an analogue database of historical records in the form of magazines and books. As internet usage became more prevalent in the 1990s and early 2000s, Ancestry shifted its focus to digitizing its records and expanding its reach beyond the United States. The subscription-based Ancestry.com not only allowed its customers to access various pieces of historical data, but it became a social network on which users could collaboratively construct their family trees, connect with distant relatives, and consult professional genealogists. The company launched its first iteration of at-home autosomal DNA tests in 2012 and over the next decade, the popularity of the product skyrocketed: By 2022, AncestryDNA had received approximately 21 million samples from consumers around the world (Jiang, Liberti, & Lebo, 2023).

Autosomal genetic testing is a relatively new technology, but it is particularly powerful in that it can identify genetic material from both the paternal and maternal lines (Paull & Akaha, 2012). This process entails the analysis of individual nucleotides (the molecular basis of DNA) within each genome, the complete set of genetic material stored in a cell's chromosomal structures (National Human Genome Research Institute, 2023). A process called phasing is used to isolate the genetic components inherited from each parent. At

this point, a supervised machine learning algorithm such as a hidden Markov model may be deployed to iterate over segments of the DNA and infer its geographic origin by comparing it with reference samples of known origin. Since the use of hidden Markov models is well documented for various tasks in bioinformatics including gene prediction and sequence comparison (Schuster-Böckler & Bateman, 2007), the present audit will not focus on this aspect of the AncestryDNA Ethnicity Estimate. A comprehensive review of the development, refinement and validation of the reference panel data is presented, instead.

Having a robust and representative dataset of the genetic traits common to many different locations is imperative to generating an accurate approximation of a customer's genetic origins. Adrion et al. (2023) outlines the process of curating such a dataset in the company's most recent Ethnicity Estimate white paper. The first step is identifying candidates with appropriate genetic profiles. This is also the root of a significant methodological limitation, as historical samples would be needed to properly map modern customer profiles to those of their potential ancestors. Instead, the company leverages its access to their database of family trees and gains consent to utilize genetic data from individuals who have been able to consistently trace their ancestry to a specific location. To account for historical factors (and colonial legacies) that have contributed to admixture among indigenous populations of the Americas and Pacific Islands, some candidates who identify with these populations are selected, despite not necessarily being able to reference longstanding historical ties to a single area. The samples from these candidates are processed such that they only contain pieces of the genome that are matched to corresponding indigenous groups by previous iterations of the Ethnicity Estimate (Adrion et al., 2023). Information about how these previous models inferred indigenous ancestry does not seem to be publicly available among the series of AncestryDNA white papers. For increased population diversity, additional samples are added to the panel from public research databases including the Human Genome Diversity Project, 1000 Genomes Project and the Human Origins dataset.

The reference panel data is refined by removing samples that share significantly long

haplotypes or stretches of genetic material. This is necessary to properly infer the probability that a given haplotype is most similar to a regional group in the reference panel. Genetic segments shared between close relatives, descended from a common ancestor within the last 10 generations, can bias the representation of a population, thus causing a threat to the validity of the haplotype matching (Ball et al., 2018; Adrion et al., 2023). Furthermore, candidates with family records in the Ancestry database are vetted to ensure that their genetic profiles match their self reported pedigree; outlier samples are identified and discarded after conducting principal component analysis (PCA) on data reflecting common genomic variants or single nucleotide polymorphisms (SNPs). PCA is a dimension reduction technique, often used in population genetics to isolate key structural elements of a dataset by projecting the data onto a plane with orthogonal axes representing linear combinations of SNP markers, in the directions of maximal variation (McVean, 2009). AncestryDNA analyses a series of two-dimensional scatter plots using the first four principal components (the four dimensions, most influential to variation) as axes, showcasing candidates' sample data. The samples are colour coded by country of origin, and it is most often observed that points corresponding to samples from geographically proximal populations form distinct clusters (Adrion et al., 2023). Outliers are defined as candidates whose samples are found to be relatively distant from those associated with their reported geographic lineage or much closer to a different group. As the reference database increases in size and diversity, areas of overlap or subgroups may become visible, allowing for more precision and granularity of the encoded biogeographical categorizations. PCA is conducted multiple times on different scales of data, narrowing from all samples across continents to those from country-level populations. Lastly, the cluster designations from the PCA are cross validated, by testing the accuracy of the hidden Markov model, trained on 95 percent of the data in the reference panel, in imputing the ethnic group labels for the remainder of the samples. Training and test sets are repeatedly shuffled, and the groupings are further refined to maximize the average predicted ethnicity accuracy for the samples from each region (Adrion et al., 2023). The current AncestryDNA reference panel consists of over 70,000 samples from 88 regional groups.

1.2 Motivation for Interrogating the AncestryDNA Reference Panel Clustering Process

The purpose of this report is to present a qualitative audit of the ideologies, concepts and statistical methodologies that form the basis of the ethnogeographic clusters produced in the AncestryDNA Ethnicity Estimate. Although the company does not explicitly use the term "race," the association of geography with ethnic community membership and biological features has its roots in anachronistic pseudoscience and the colonial interests that led to the social construction of race (Belen, 2018). This audit does not attempt to claim that Ancestry or similar companies are actively pursuing a racist agenda, however it does endeavour to show that systems like the Ethnicity Estimate produce racialized profiles, which have been historically used to keep structures of oppression intact.

On the other hand, the production and refinement of machine learning based ethnicity estimates serves to fulfil an existing consumer demand. As suggested by the tens of millions of AncestryDNA test kits sold between 2012 and 2022, there is an increasing trend of public interest in genetic and genomic testing. This interest is stoked by advertising slogans encouraging potential customers to "Discover the [stories] only [their] DNA can tell" (Peters, 2023). Such statements not only obscure the statistical uncertainty and complex nature of the Ethnicity Estimates, but they also attract those with vested personal desires to identify with a community, connect with long-lost biological relatives or engage in other means of self discovery (Bolnick et al., 2007). Given its profit driven nature, the company capitalizes on the rhetorical notion that consumers will only truly be able to understand who they are with the help of cost prohibitive subscriptions, genealogy consulting services or ancestry tests.

The following section presents a comprehensive review of scholarly literature focusing on the past scientific construction of racialized ethnogeographic groups, as well as an overview of qualitative and data driven audits of DTC genetic testing services. Then a qualitative audit of the Ancestry Ethnicity Estimate with a focus on reference panel clusters is developed using the Algorithmic Ecology framework proposed by the Stop LAPD Spying Coalition and Free Radicals (2020).

2 Literature Review

2.1 Early Statistics, Datafication and the Modern Construction of Race

Past research in algorithmic fairness, accountability and transparency has developed the notion of "datafication," in reference to the practice of organizing and often simplifying nuanced concepts or phenomena such that they are easily catalogued, quantified and analysed (Valdivia & Tazzioli, 2023). Although the word, data, evokes thoughts of modern computers and large scale algorithms, datafication lies at the basis of many centuries of scientific inquiry. In his 2006 essay, "Making Up People," Ian Hacking alludes to this idea in the context of the social or human sciences—He discusses how people are sectioned off into groups such that their problems may be addressed using empirical methods, and that the understanding of these groupings leads their members to interact with the world in new (made up) ways. The categorizations are almost always arbitrary, reflecting the bias of scientists and institutions who have the power to decide who is and is not classified as "normal," as well as determine hierarchical social structures dictating patterns of acceptance and stigma (Hacking, 2006).

Racial categories are a prime example of the legacy of historical datafication and making up people. While there is empirical evidence undermining race as a biological construct (McCann-Mortimer, Augoustinos, & LeCouteur, 2004), the origin of the phenomenon has led to severe social consequences that continue to impact humanity. Carl Linnaeus is broadly cited as the founder of modern biological taxonomy. In 1758, he theorized that there were distinct human races native to the continents of Africa, Europe, Asia and the Americas (Belen, 2018). Natural scientists and anthropologists continued to build upon this work for the next two centuries, searching for phenotypic markers of racial difference from skull-shapes to fingerprint patterns. Francis Galton, a pioneer of both modern statistics and eugenics, popularized the use of biometric data as a means of tracking, surveilling and punishing racialized subjects under 19th century British colonial rule (Valdivia & Tazzioli, 2023).

Datafication has only scaled up into the modern age, as seen in the following sections.

2.2 Data-Driven Audits of DTC Genetic Ancestry Test Results

"Keeping Data Alive"

Direct algorithmic audits of DTC genetic testing companies have been severely limited by company data policies and the proprietary nature of commercial technologies. Nonetheless, various scholars have developed creative angles for interrogating these companies' business practices (Bolnick et al., 2007), biomedical contributions (Tandy-Connor et al., 2018) and, of course, social implications. A notable contribution to this body of critique is Ruckenstein (2017)'s discussion of how personal data created by DTC genetic test results enhances knowledge production and identity related discourse among users. Focusing on 23andMe, Ruckenstein introduces the "lively data" framework, which seeks to observe how this personal data is kept alive not only through its impact upon how subjects view their identity narratives, but also how DTC testing companies repurpose genetic data for the refinement of future prediction models. Additional consideration is given to data that is ultimately forgotten as insignificant to consumers or unusable for companies. In a similar vein, Nafus (2014) states "...the question of which data might become part of what kinds of social entanglements, and which data will remain indecipherable and inconsequential, becomes clear only with hindsight."

The notion of "lively data" is operationalized by tracking the ways in which 23andMe customers share information about their genetic test results, with the primary goal of observing and recording patterns in "gene talk" that contribute to mainstream understandings of population genetics and opinions regarding the validity of the tests. Ruckenstein (2017) draws from previous work to assert that 23andMe and other genetic ancestry testing companies create an inhabitable map, "...a pleasurable way of wandering in the world that is simultaneously a poignant reminder of how we become observed." With the help of interactive user interfaces and visualizations, consumers are made to believe that they have agency over their genetic identity profiles, despite the near total lack of transparency in terms of how these results are derived. Essentially, the inhabitable map is

strikingly similar to the datafication carried out by colonial scientists like Francis Galton; however, the former frames the subjects of datafication as having a semblance of power, comparable to the institutions that produce and profit from their identities.

To further investigate these dynamics, a qualitative study is outlined, which entails an iterative process of participatory research. Twenty Finnish adults between the ages of 22 and 55 were recruited to engage in surveys and discussions about their use of 23andMe, as well as contribute thoughts and feedback on others' comments at later stages. The primary findings indicate that participants' "gene talk" consisted of alternative lenses through which to interpret their test results; those who did not blindly accept the significance of their places in the inhabitable map tended to direct more focus towards understanding how the test results could be translated into meaningful information. For example, many of the participants were disappointed to hear that most (upwards of 90 percent) of their genetic lineage came from Finland; one even remarked that such results could have just as easily been deduced from "his address and credit card information" as they could have from his DNA (Ruckenstein, 2017). In this case, the generic nature of the results showed that DTC ancestry testing companies were not fulfilling their promise of telling the stories that only one's DNA can reveal. Some participants bonded over the novelty of having reported Neanderthal genetic input in their test results and being able to trace prehistoric roots. Beyond the desire for exciting or unexpected outcomes, other participants sought to transpose their positions within 23andMe's inhabitable map onto astrological or religious mappings that aligned more with their personal beliefs, thus creating more meaningful interpretations of the data.

Promethease

Ruckenstein (2017) also notes a general interest among participants in accessing their raw genetic data. The ability to simply access their samples made many participants feel as though they could play a role in contributing to scholarship in population genetics and/or medicine, either by donating their data to institutions in these fields or conducting their own analysis—though there remains a high barrier to entry in terms of obtaining complex

scientific domain knowledge or seeking advice from professional genetic researchers. Some participants were also aware of the Promethease platform and took advantage of it as a tool to double-check the validity of their 23andMe results.

Developed by geneticist, Greg Lennon, and bioinformatician, Mike Cariaso, Promethease is a free downloadable software that allows users to upload raw genetic data (often obtained from DTC testing profiles such as AncestryDNA and 23andMe) and receive more detailed reports on their local machines (Cariaso & Lennon, 2012). These reports contain information regarding the frequency of their genetic variants compared to sample populations, and an estimated measure of the impact of these variants in terms of their medical, phenotypic and genealogical associations (Hayden, 2008). Promethease pulls genetic records from a wiki SNPedia, which contains data on over 100,000 SNPs; this material is collected through routine searches of public databases (e.g. dbSNP, HapMap, Ensembl, and PharmGKB), automated web scraping of PubMed articles (containing unique indices or RS numbers associated with individual variants or SNPs), and voluntary user submission of genetic data and self-reported phenotypes (Cariaso & Lennon, 2012). The diversity of data types and formats on SNPedia means that Promethease reports are often contextualized with qualitative interpretations of an SNP's influence on health risks and direct hyperlinks to primary sources.

While the aim of Promethease is to make analysing genetic data more accessible, namely to casual consumers of DTC tests, the reports remain complex and somewhat difficult to decipher for a general audience. Unlike AncestryDNA or 23andMe, the platform does not function as an inhabitable map: Its interface is simple, without eye-catching graphs and easily-digestible infographics, and instead of clear-cut genetic breakdowns, it often links to more nuanced findings. Ruckenstein (2017) remarks, "From the user perspective, the report is a compilation of apparently random facts rather than a coherent risk analysis framework, mirroring the current state of scientific research that offers conflicting and uncertain information about the genetic disposition of health." Using Promethease to interrogate AncestryDNA or 23andMe test results may reasonably prompt increased scepticism of the reliability of these services, as it highlights the lack of transparency in

the processes that underlie the creation of a customer's profile. In the case of genetic variations with empirically disputed links to different populations or health conditions, how do these companies transform customer data into clear predictions or risk assessments? For some, this may seem like a dead-end in their search for self discovery on a genetic level. However, seeing the fuller scope of the results offered by a site like Prometheease does provide some level of agency in that it frames DTC tests as only one possible source of information, instead of an absolute and irrefutable classification (Hayden, 2008; Ruckenstein, 2017).

Algorithmic Ecology

As outlined in the previous subsections, past audits have called into question the ways in which DTC genetic testing companies commodify, manipulate and present customers' data, while also uplifting alternative interpretations of test results and the pursuit of external knowledge in narrative, qualitative and quantitative forms. Returning to Ruckenstein's conception of "lively data," it is clear that services like 23andMe or AncestryDNA filter out the values they deem worthy of keeping alive, though these judgments are not always consistent with the priorities of their customers. Generating a report in Prometheease revives additional data in the form of empirical discourse and uncertainty in the interpretation of specific genetic variants. The audit presented in the following section seeks to bring to life the experiences of racialized communities that have been impacted by the ethnogeographic profiles, assigned to them through vehicles such as the AncestryDNA Ethnicity Estimate.

This content lends itself nicely to the Algorithmic Ecology auditing strategy. This method was first outlined in a 2020 Medium article authored by the Stop LAPD Spying Coalition and Free Radicals, highlighting the destructive nature of the Predpol predictive policing algorithm.

> The Algorithmic Ecology is both a framework and an organizing tool that can be critically applied to any algorithm. This model decenters the algorithm itself, looks at the different actors that shape the algorithm, and illustrates

> whose interests the algorithm serves, with the ultimate goal of dismantling the actors creating algorithmic harm (Stop LAPD Spying Coalition and Free Radicals, 2020).

The four levels at which an algorithm is critiqued through Algorithmic Ecology include the community level, the operational level, the institutional level and the ideological level. All together, these components serve to contextualize the functioning of the algorithm in terms of how it affects different stakeholders.

3 The Algorithmic Ecology of AncestryDNA's Ethnicity Estimate

3.1 Community

The community level of an Algorithmic Ecology considers the people and groups who are most impacted by a technology as consumers or recipients of an intervention. In the case of the AncestryDNA Ethnicity Estimate, the community level corresponds directly to the cluster groupings in the reference panel with respect to their use for predicting the group membership of others. This may be explored in terms of how customers navigate the labels assigned to them by the Ethnicity Estimate and the factors shaping their dispositions towards the results.

Foeman (2012) showcases some of these community implications by conducting a qualitative analysis of the relationship between the ancestry profiles produced by DTC genetic tests, family narratives and the social construction of race. Although the project is framed as a class activity for undergraduate students in fields related to intercultural communication, its findings reveal more general insights about how test results may elicit sentiments reflecting hierarchical differentials in desirability or social capital associated with racialized labels. The work is grounded in the idea that individuals, families and community groups are agents in the creation and performance of narrative identities (Ricoeur, 1991). Essentially, people select the stories they tell about themselves such that they portray

a certain image. Despite their probabilistic uncertainty and oversimplification, Foeman suggests that DTC genetic testing profiles may serve as a constructive means of auditing consumers' identity narratives. In some cases, unanticipated results may reflect "the ways that narratives change to favour the survival of some racial identities over others... [constituting] more evidence of attitudes held about the value and status of certain racial associations" (Foeman, 2012).

Foeman specifically notes the reaction of Black students to ancestry results indicating European heritage as compared to White students who were shown to have some African heritage. The DNA tests served to validate the pre-interviews of most Black students, who reported having some White relatives either as a result of colonial violence or interracial relationships, whereas almost all the White participants with African ancestry found it surprising. These results are interpreted as follows:

> Despite a desire on the part of majority members to eradicate the narratives that link Black and White, the connection is preserved in African American recollections and DNA. Maintaining a racial identity that includes a European element, still identified as the culture of power, beauty, and prestige (Jackson, 1999; Mcintosh, 1990), is desirable in spite of a negative past and because of it (Foeman, 2012).

Although the exact brand of the DNA testing kits used in this project is not revealed, Foeman's work still suggests that these tests are generally viewed through a racialized lens—even though DTC genetic test companies strategically use language relating to ethnicity and region to avoid the explicit mention of race (Chomsky, 2018; Girard, 2020; Peters, 2023). It may be argued that, whether completely accurate or not, these kinds of test results force users to confront their subconscious notions of racial hierarchies (e.g. reflected in a desire for proximity to Whiteness or "pure" European ancestry); however, the Ethnicity Estimate (or any other ancestry test service) does little to actively invite this kind of reflection, let alone disrupt such notions of hierarchy.

3.2 Operational

The operational level of an Algorithm Ecology relates to how potential harms are operationalized by a given technology. This often entails how notions of algorithmic objectivity or value neutrality are reinforced by the companies that implement the technology. Furthermore, the operational level tends to exemplify a lack of transparency with respect to the entire scope of an algorithm's function.

Girard (2020) discusses how companies such as AncestryDNA have capitalized on the use of big data and machine learning technologies, both in terms of reducing the cost and human labour necessary for conducting large-scale genetic research and in selling the belief in the objective authority of their predictions; she writes: "Profit maximization, in short, is rebranded as bias minimization." Nonetheless, there is still a level of human bias in determining and updating the regional samples in the AncestryDNA reference panel. There is a level of subjectivity to the clustering process as PCA results are "visually inspected" (Adrion et al., 2023), presumably by human researchers. Identifying an outlier in a scatter plot visualization may seem intuitive, however, the company's most recent white papers do not describe any protocol for consistently determining which samples are too distant from their respective clusters and when overlapping regional groups constitute a new subgroup (or should be consolidated). The 2023 white paper briefly mentions the use of an additional dimension reduction technique, Uniform Manifold Approximation and Projection, which does more explicitly account for the distance between points and clusters when transposing them onto two dimensions (Diaz-Papkovich, Anderson-Trocmé, Ben-Eghan, & Gravel, 2019). However, the Ancestry paper does not provide any information about the method or the cases in which it is used, e.g. on local or global scale data.

Changes in the quantity and organization of reference panel groups may lead to regional categories being added, swapped or completely removed from a user's profile (Girard, 2020). According to the AncestryDNA web page titled "Ethnicity Updates," current customers will receive free profile updates with each new iteration of the Ethnicity Estimate and their prior results will only be available for 90 days following the new release (Ancestty Support, 2024) A continually improving product renders changes

in customers' identity profiles out of their control.

3.3 Institutional

Institution-level considerations within an Algorithmic Ecology concern the specific entities that back the implementation of a technology, with a vested interest in its social implications. For private companies such as Ancestry, this may be a difficult theme to investigate, as the company likely conflates the functioning of its products with profit, more than any other institutional incentive. Nonetheless, the action of operating a profit based model within contexts that could potentially harm consumers or undermine their communities must be examined.

Peters (2023) frames DTC ancestry testing as operating within the institution of western science, which has historically served to validate colonial interests and ideologies such as white supremacy (Seth, 2009; DiGangi & Bethard, 2021). Whereas the use of state-of-the-art technologies in interdisciplinary fields such as bioinformatics is often discussed in terms of being disruptive, it is important to acknowledge the specific nature of their disruption: It unsettles at the level of communities and individuals, but not the institutions that inform social and power structures (Chun, 2021). For example, Peters (2023) highlights the fact that some of AncestryDNA's ethnogeographic groupings reference Bantu people—a linguistic classification, coined by the German philologist, Wilhelm Bleek, whose work emphasized the group's primitiveness (Bank, 2000). Although Africans and Black diasporic communities have previously rejected the term as an ethnic label, AncestryDNA's revival of the label in this context symbolizes an assertion of authority over how Black (often Black American) customers identify themselves in relation to a broader African diaspora.

3.4 Ideological

Lastly, the ideological level of the Algorithm Ecology seeks to break down and critique the hegemonic value systems upheld by a given technology. As previously mentioned, the development of racialized classifications is the result of scientific practices rooted in western

superiority, white supremacy and colonialism. One ideological aspect of these institutions is rooted in the historical extermination of indigenous people and the subsequent notion that these groups do not exist in modernity. AncestryDNA tacitly contributes to this ideology by assuming that Native reference populations are inherently admixed (Adrion et al., 2023) and that most people's connections to indigenous identity are extremely distant, if not completely lost to biological dilution (Chomsky, 2018; Girard, 2020). Such assumptions can have lasting implications for the civil and human rights of indigenous people.

Blood quantum is a concept rooted in eugenics that still plays a role in determining who is eligible for tribal citizenship in the United States. The Indian Reorganization Act passed in 1934 legally defines Native Americans as either members of a federally recognized tribe, descendants of tribal citizens who were living on reservations as of June 1, 1934, and anyone else who can prove that they have at least 50 percent Native American ancestry or "blood" (Schmidt, 2011). Although blood quantum is still recorded by many tribal sovereignties, it comes out of fears relating to histories of genocide at the hands of the US government. Not accounting for blood quantum may increase poverty due to scarce resources resulting from increased populations as well as provide opportunities for outsiders to enter positions of power within tribal governments and "sell the land" (Native Governance Center, 2024). On the other hand, continuing to identify Native Americans in terms of blood quantum could serve to completely eliminate them from the recorded population. Furthering this point, Limerick (as cited in Schmidt, 2011) writes:

> Set the blood quantum at one-quarter, hold to it as a rigid definition of Indians, let intermarriage proceed as it had for centuries, and eventually Indians will be defined out of existence. When that happens, the federal government will be freed of its persistent "Indian problem."

The ways in which DNA test companies simplify the reporting and calculation of ancestry upholds a history of datafication, with the seeming goal of making the "vanishing Indian" stereotype (Berry, 1960; Chomsky, 2018) a reality.

4 Conclusion

To summarize, this work presents a critical analysis of the AncestryDNA Ethnicity Estimate system and other DTC genetic testing services through the lens of the Algorithmic Ecology Framework. It is evident that these technologies have varied implications for racialized communities and, in many ways, carry on the oppressive legacies of datafication and the social-scientific construction of race. This audit only begins to unpack the experiences of those whose identities have been shaped by genetic ancestry test results. Future scholarship should seek to investigate individual narratives as well as effects on other ethnic groups or diasporas. Further consideration should also be given to platforms like Promethease, if genetic data may be ethically obtained, as well as Ancestry's social network and archival resources. Since the demand for DTC genetic tests only continues to grow, resources should be produced such that consumers are empowered with relevant and contextualized information to construct, reflect upon and ultimately take ownership of their identity narratives.

References

Adrion, J., Lang, J., Noto, K., Sedghifar, A., Olpin, R., Wang, Y., & Wolf, A. (2023). Ethnicity estimate 2023 white paper. Retrieved from `https://support.ancestry.com/s/article/AncestryDNA-White-Papers?language=en_US`

Ancestty Support. (2024). *Ancestrydna® ethnicity updates.* https://support.ancestry.com/s/article/AncestryDNA-Ethnicity-Updates?language=en$_US$.

Ball, C., Battat, E., Byrnes, J. K., Carbonetto, P., Chahine, K. G., Curtis, R. E., ... others (2018). Genetic communities™ white paper: Predicting fine-scale ancestral origins from the genetic sharing patterns among millions of individuals. *Ancestry.[https://www. ancestry. com/cs/dna-help/communities/whitepaper]*.

Bank, A. (2000). Evolution and racial theory: the hidden side of wilhelm bleek. *South African historical journal*, *43*(1), 163–178.

Belen, D. (2018). How cranial shapes led to contemporary ethnic classification: a historical view. *Turkish Neurosurgery*, *28*(3), 490–494.

Berry, B. (1960). The myth of the vanishing indian. *Phylon (1960-)*, *21*(1), 51–57.

Bolnick, D. A., Fullwiley, D., Duster, T., Cooper, R. S., Fujimura, J. H., Kahn, J., ... others (2007). The science and business of genetic ancestry testing. *Science*, *318*(5849), 399–400.

Cariaso, M., & Lennon, G. (2012). Snpedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic acids research*, *40*(D1), D1308–D1312.

Chomsky, A. (2018, November 29). Dna tests make native americans strangers in their own land. *The Nation*. Retrieved from `https://www.thenation.com/article/archive/dna-tests-elizabeth-warren-native-american-race-science/`

Chun, W. H. K. (2021). Discriminating data. *Filmed on September*, *28*.

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., & Gravel, S. (2019). Umap reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*, *15*(11), e1008432.

DiGangi, E. A., & Bethard, J. D. (2021). Uncloaking a lost cause: Decolonizing ancestry

estimation in the united states. *American Journal of Physical Anthropology*, *175*(2), 422–436.

Foeman, A. K. (2012). An intercultural project exploring the relationship among dna ancestry profiles, family narrative, and the social construction of race. *Journal of Negro Education*, *81*(4), 307–318.

Girard, A. K. (2020). Algorithms, identity, and cultural consequences of genetic profiles. In *Algorithmic culture: How big data and artificial intelligence are transforming everyday life* (p. 125-139). Lexington Books.

Hacking, I. (2006). Making up people. *London Review of Books*, *28*(16), 23-26.

Hayden, E. C. (2008). How to get the most from a gene test. *Nature*, *456*, 6.

Jiang, S., Liberti, L., & Lebo, D. (2023). Direct-to-consumer genetic testing: A comprehensive review. *Therapeutic Innovation & Regulatory Science*, *57*(6), 1190–1198.

McCann-Mortimer, P., Augoustinos, M., & LeCouteur, A. (2004). 'race'and the human genome project: constructions of scientific legitimacy. *Discourse & Society*, *15*(4), 409–432.

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS genetics*, *5*(10), e1000686.

Nafus, D. (2014). Stuck data, dead data, and disloyal data: the stops and starts in making numbers into social practices. *Distinktion: Scandinavian Journal of Social Theory*, *15*(2), 208–222.

National Human Genome Research Institute. (2023). *Human genomic variation.* https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation.

Native Governance Center. (2024). *Blood quantum and sovereignty: A guide.* https://nativegov.org/resources/blood-quantum-and-sovereignty-a-guide/.

Paull, J. M., & Akaha, J. B. (2012). Using autosomal dna analysis to connect rabbinical lineages: A case study of the wertheimer and wertheim dynasties. *AVOTAYNU: The International Review of Jewish Genealogy*, *28*(4).

Peters, C. (2023). Racial-genomic interest convergence and the geneticization of black

families. *Journal of Family Communication*, *23*(3-4), 294–309.

Ricoeur, P. (1991). Narrative identity. *Philosophy today*, *35*(1), 73.

Ruckenstein, M. (2017). Keeping data alive: Talking dtc genetic testing. *Information, Communication & Society*, *20*(7), 1024–1039.

Schmidt, R. W. (2011). American indian identity and blood quantum in the 21st century: A critical review. *Journal of Anthropology*, *2011*.

Schuster-Böckler, B., & Bateman, A. (2007). An introduction to hidden markov models. *Current protocols in bioinformatics*, *18*(1), A–3A.

Seth, S. (2009). Putting knowledge in its place: science, colonialism, and the postcolonial. *Postcolonial studies*, *12*(4), 373–388.

Stop LAPD Spying Coalition and Free Radicals. (2020). *The Algorithmic Ecology: An Abolitionist Tool for Organizing against Algorithms.* Medium. Retrieved from `https://stoplapdspying.medium.com/the-algorithmic-ecology-an-abolitionist-tool-for-organizing-against-algorithms-14fcbd0e64d0`

Tandy-Connor, S., Guiltinan, J., Krempely, K., LaDuca, H., Reineke, P., Gutierrez, S., ... Davis, B. T. (2018). False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. *Genetics in Medicine*, *20*(12), 1515–1521.

Valdivia, A., & Tazzioli, M. (2023). Datafication genealogies beyond algorithmic fairness: making up racialised subjects. In *Proceedings of the 2023 acm conference on fairness, accountability, and transparency* (pp. 840–850).