

5-18-2024

DIGITAL ALLOTMENT AND VANISHING INDIANS: IDSOV AND LLMS

Sam McVeety
University of Washington

Follow this and additional works at: <https://digitalcommons.law.seattleu.edu/ailj>



Part of the [Administrative Law Commons](#), [Civil Rights and Discrimination Commons](#), [Criminal Law Commons](#), [Energy and Utilities Law Commons](#), [Environmental Law Commons](#), [Gaming Law Commons](#), [Indigenous, Indian, and Aboriginal Law Commons](#), [Intellectual Property Law Commons](#), [Land Use Law Commons](#), [Law and Gender Commons](#), [Transnational Law Commons](#), and the [Water Law Commons](#)

Recommended Citation

McVeety, Sam (2024) "DIGITAL ALLOTMENT AND VANISHING INDIANS: IDSOV AND LLMS," *American Indian Law Journal*: Vol. 12: Iss. 2, Article 4.

Available at: <https://digitalcommons.law.seattleu.edu/ailj/vol12/iss2/4>

This Article is brought to you for free and open access by the Student Publications and Programs at Seattle University School of Law Digital Commons. It has been accepted for inclusion in American Indian Law Journal by an authorized editor of Seattle University School of Law Digital Commons. For more information, please contact coteconor@seattleu.edu.

DIGITAL ALLOTMENT AND VANISHING INDIANS: IDSOV AND LLMs

Sam McVeety¹

The rapid proliferation of Large Language Models (LLMs) like ChatGPT and Google's Bard represents a new variation on an old theme in Federal Indian law: the imposition of a monolithic worldview with a false veneer of universality. From Johnson v. M'Intosh onwards, federal Indian law is characterized by non-Native actors creating rigid definitions of what it is and how to be Indigenous in the area now called the United States and reifying those definitions through legal precedent and state violence.

Having been tuned to maximize apparent credibility, LLMs similarly co-create reality, by offering ostensibly objective statements on reality. Just as legal precedents and state violence in federal Indian law creates self-reinforcing feedback loops regarding culture and societal fitness, so too do LLMs engender a self-reinforcing, colonially inflected worldview.

In this Article, I examine the connections between Federal Indian law, LLMs, Indigenous Sovereignty Data (IDSov), and the patterns that reappear across these domains. Current data gathering practices stand to amplify harm to Native interests, and Native Nations currently face difficult choices between perpetuating their own erasure or accepting gross misrepresentations. Although there are severe shortcomings in existing legal protections, there are potential legal and technical solutions to deploy that are rooted in tribal sovereignty.

I. INTRODUCTION

Representation matters. Being here matters. No longer do Indigenous people have to occupy a site to get the attention of the government. We are at the table and we are not giving up our seats.²

At the core of identity is the power to say who you are *and* the power to be seen that way by others. In being seen by others, there are two dimensions: how you live and how structures of

¹ As a Master of Jurisprudence student at the University of Washington, this work represents the intersection of my own legal studies and technological expertise. In writing this paper, I would like to thank Professors Monte Mills, Rebecca Tsosie, Angélica Cházaro, Huy Nguyen, and Theodore Myhre; Deans Mario Barnes and William Covington; Christina Greer, Salem Haykal, Jevan Hutson, Eric Brewer, and Jordan Goldwarg.

² Debra A. Haaland, *Secretary Haaland Delivers Remarks on the Power of Indigenous Representation on the 52nd Anniversary of the Alcatraz Island Occupation*, U.S. DEPARTMENT OF THE INTERIOR, OFFICE OF THE SECRETARY (Sept. 9, 2023), <https://www.doi.gov/pressreleases/secretary-haaland-delivers-remarks-power-indigenous-representation-52nd-anniversary>.

power categorize you. From the very beginnings of colonial contact on Turtle Island,³ colonial actors have contested Indigenous peoples' power to be seen.⁴ Outside narratives of how Native Nations lived informed how the law and state power regarded them, turning subjective differences into objective harm. Through boarding schools,⁵ the Indian Reorganization Act (IRA), model constitutions,⁶ blood-quantum-based allotments,⁷ and Termination Era policies,⁸ the United States government has tried to violently proscribe this power to be seen, telling Indigenous peoples who they are, what forms of government they are allowed, how "Indian" they are, and whether they are a Native Nation at all.⁹

The United States legal system and adjacent fields of data gathering co-construct reality in specific, settler-colonial inflected ways. Concepts like blood quantum begin as legal constructs, which then motivate data gathering exercises (e.g. the decennial census), which then inform supposedly objective "facts" that future legal doctrines draw upon (e.g. the "number of

³ See ROBIN W. KIMMERER, *BRAIDING SWEETGRASS*, 4-10 (Milkweed Editions, 2013) (now also known as North America).

⁴ Throughout, I will attempt to use terminology thoughtfully. As a general rule, I will use "Native Nation" rather than "tribe" to refer to sovereign groups, although I will use "tribal" to refer to concerns relating to those groups (e.g. "tribal sovereignty"). In keeping with the focus on IDSov, I will primarily use "Indigenous" when describing adjacent concepts, and when describing a more general frame than the United States. In selecting language, Gregory Younging notes that context differs between the United States and Canada. Gregory Younging, *Elements of Indigenous Style: A Guide for Writing by and About Indigenous Peoples*. (Edmonton Alberta: Brush Education, 2018). Here, I will borrow an approach from David Treuer: "Throughout ..., I use the word "Indian" to refer to Indigenous people within the United States. I also use "Indigenous," "Native," and "American Indian." These terms have come in and out of favor over the years, and different tribes, not to mention different people, have different preferences. The Red Lake Nation refers to itself as the "Home of the Red Lake Band of Chippewa Indians," for example. Many Native people prefer to describe themselves in their Native languages: Piikuni for Blackfeet, Ojibwe for Chippewa, and so on. My own choices of usage are governed by a desire for economy, speed, flow, and verisimilitude. A good rule of thumb for outsiders: Ask the Native people you're talking to what they prefer." DAVID TREUER, *THE HEARTBEAT OF WOUNDED KNEE: NATIVE AMERICA FROM 1890 TO THE PRESENT 1* (Riverhead Books 2019).

⁵ In a devastating reification of "kill the Indian, save the man," federal government agents and non-governmental organizations forcibly removed children from their families in order to reeducate them in white-dominant culture and strip them of Indigenous identity and culture. See Denise K. Lajimodiere, *STRINGING ROSARIES: THE HISTORY, THE UNFORGIVABLE, AND THE HEALING OF NORTHERN PLAINS AMERICAN INDIAN BOARDING SCHOOL SURVIVORS* (University of Manitoba Press 2019).

⁶ While the Indian Reorganization Act of 1934 created some protections for Native Nations, it also created vast changes in federal Indian law without explicit tribal consent. A particular feature of the IRA was the model constitution that it pushed Native Nations to adopt, superseding and erasing historical practices of self-governance. Indian Reorganization Act of 1934, ch. 576, 48 Stat. 984 (codified as amended at 25 U.S.C. §§ 461–479).

⁷ During the allotment era, during which the federal government stole millions of acres of land from Native Nations, the federal government judged the "Indian-ness" of allottees by the amount of Indian lineage in their genetic history. General Allotment Act of 1887, ch. 119, 24 Stat. 388 (codified as amended at 25 U.S.C. §§ 331–358).

⁸ TREUER, *supra* note 4, at 231-280. During the Termination Era, the federal government called the very legitimacy of Native Nations into question, demanding that Native Nations justify their existence (implicitly with respect to the white gaze) through a series of bureaucratic hurdles defined without the explicit consent of Native Nations. The federal government ceased to recognize ("terminated") Native Nations that could not meet these standards, removing their access to various benefits and protections.

⁹ *Id.* at 378-380. The federal government deployed blood quantum as a means of determining whether individuals were Indian or not. The IRA imposed a specific system of government on Native Nations. Indian Reorganization Act, *supra*. Federal policy during the Termination Era declared that some Native Nations no longer existed. The practice of requiring federal recognition for Native Nations continues to this day.

Indians in existence").¹⁰ This feedback loop, across disciplines that are typically studied independently, serves to obscure the artificiality and subjectivity of this co-constructed reality. Actors in each domain (legal scholars and data analysts) take the word of the other as "truth," all the while participating in a legal-data project that supports a specific, contingent, settler-colonial worldview.¹¹

Thus, Native control of information about the Native Nation and its members is a central aspect of sovereignty.¹² The story of Federal Indian law is frequently one of toxic paternalism, with data and information (valid or not) playing a central role in justifying destructive policies across multiple centuries.¹³ Even for well-established Native Nations in the current era of tribal self-determination, the nature and content of such information can have far-reaching economic and cultural impacts. These impacts include access to capital, protection of traditional knowledge, custody of children, and access to medical care.¹⁴

Recent developments in artificial intelligence and machine learning threaten Indigenous peoples' power to define who they are in new and dangerous ways. In the past few decades, society's relationship with information has changed rapidly, through the emergence of "big data" practices and widespread and massively scalable analytical tools.¹⁵ One of the latest developments in this field is the emergence of LLMs, massive systems which can mimic many human response patterns. This draws on a discipline of machine learning that abstracts actor and action within a monolithic set of computer instructions, drawing inferences from huge datasets. The creation, development, and deployment of these models pose a direct threat to Indigenous self-representation, by perpetuating mis- and underrepresentation of Indigenous interests, and co-opting Indigenous cultural property.

Although Native Nations have made strides in recent years to codify defensive data practices within the field of IDSov,¹⁶ the collision of these emerging safeguards and LLMs

¹⁰ See *infra* note 37. Described by Maggie Walter as the "data of disregard."

¹¹ This is exemplified by Justice Rehnquist's "excus[ing of] the government's grievous mistreatment of the Lakota in the late nineteenth century in his dissenting opinion in *United States v. Sioux Nation of Indians*," based on a "both sides" view of history: "Undoubtedly greed, cupidity, and other less-than admirable tactics employed by the Government during the Black Hills episode in the settlement of the West, but the Indians did not lack their share of villainy either." Rebecca A. Tsosie, *Reclaiming Native Stories: An Essay on Cultural Appropriation and Cultural Rights*, 34 ARIZ. ST. L.J. 320 (2002), <https://ssrn.com/abstract=1401522>.

¹² From across an ocean, the doctrine of discovery legally wrested away (in the courts of Europe) the ability for Native Nations to call themselves sovereign and assert that their territories were not "uninhabited." The United States Supreme Court endorsed this thinking in *Johnson & Graham's Lessee v. McIntosh*, 21 U.S. 543 (1823).

¹³ See TREUER, *supra* note 4. Efforts like allotment often adopted the guise of "concern" for the Indian. Also, "[b]elieving that selfishness was essential to civilization, Dawes did his best to impose it upon tribes across the country." See also Marjane Ambler, *The Long Tradition of Defying Selfishness*, 7 TRIBAL COLL. J. AM. HIGHER EDUC. (1995), <https://tribalcollegejournal.org/long-tradition-defying-selfishness/>.

¹⁴ See, e.g., *Havasupai Tribe v. Arizona Bd. of Regents*, 204 P.3d 1063 (Ariz. Ct. App. 2009); *Haaland v. Brackeen*, 599 U.S. 255 (2023).

¹⁵ See VIRGINIA EUBANKS, *AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH* (New York, NY: St. Martin's Press, 2017) (An overview and several case studies related to abuse of big data and the potential harms to marginalized communities).

¹⁶ See Rebecca Tsosie, *Tribal Data Governance and Informational Privacy: Constructing "Indigenous Data Sovereignty"*, 80 MONT. L. REV. 229 (2019) (For a comprehensive treatment of the topic).

presents an urgent, challenging field for Native Nations and tribal members.¹⁷ There is no clear path to minimizing near-term harms, and the possible solutions in this space assume complex interactions among technology, Native Nations, government, and private entities. In this paper, I will examine the problem in several parts: (1) background on IDSoV and LLMs, (2) potential harms of this technology, (3) relevant data protection law, and (4) possible solutions.

As a white author and observer, I intend to use this paper to draw attention to these issues and amplify, connect, and support the Indigenous-led work across disciplines that is already happening at these intersections. I am keenly aware of the implicit coloniality of much Indigenous scholarship—both past and present—and seek to represent others' experiences in their words wherever possible.¹⁸

II. THE NEXUS OF IDSOV, LAW AND LLMs

The history of Federal Indian law is characterized by legal actions based on incomplete information (whether inadvertently or willfully). Beginning with *Johnson v. M'Intosh*, the United States Supreme Court fashioned the foundations of Federal Indian law based on a specific property dispute, while maintaining a paternalistic view of Native perspectives and interests.¹⁹ The latter two entries in the "Marshall Trilogy" (*Cherokee Nation v. Georgia* and *Worcester v. Georgia*) expanded their ambit to deeper questions of sovereignty, inventing the formulation of "domestic dependent nations" and applying it wholesale to Native nations. In later attempts to measure "Indianness," agents of the United States government employed techniques such as blood quantum and phrenology to try to accomplish this characterization.²⁰ Throughout, Native Nations have been coerced towards framing themselves in reductive ways that render them legible to the United States government and its laws.²¹

Despite resting on incomplete bases, these legal foundations reify over time, taking socially constructed ideas and giving them import in the physical world.²² In Federal Indian law, Native Nations were displaced, robbed, and murdered based on subjective (and self-serving) interpretations of what the law allowed. Blood quantum has since become a fixture in many

¹⁷ In one recent study of such models, researchers found that the likelihood of racist responses from a model actually increased *superlinearly* to the size of a model. That is, roughly, the larger the model gets, the more extreme its racism. Abeba Birhane et al, *On Hate Scaling Laws For Data-Swamps*, ARXIV (Jun 23, 2023), <https://arxiv.org/pdf/2306.13141.pdf>.

¹⁸ Richard Delgado, *The Imperial Scholar: Reflections on a Review of Civil Rights Literature*, 132 U. PA. L. REV. 577 (1984) <https://doi.org/10.2307/3311882> ("We are talking about you, not to you.").

¹⁹ This case established that the United States government had the exclusive right (under the doctrine of discovery) to Indian title. *Johnson & Graham's Lessee v. McIntosh*, 21 U.S. 543 (1823). See also Matthew L. Fletcher, *Listen*, 3 MICH. J. OF RACE & L. 523, 530 (1998) ("Johnson v. M'Intosh is a lie. It is wrongly decided. You may not kill people and destroy what they are and call it legal and fair play.")

²⁰ See e.g., *Cherokee Nation v. Georgia*, 30 U.S. 1 (1831), *Worcester v. Georgia*, 31 U.S. 515 (1832).

²¹ See TREUER, *supra* note 4, at 146.

²² See IAN F. HANEY LÓPEZ, *WHITE BY LAW: THE LEGAL CONSTRUCTION OF RACE*, 10TH ANNIVERSARY EDITION (New York University Press 2006) (For a thorough exploration of the legal construction of race).

tribal constitutions, dictating eligibility for membership and benefits.²³ Thus, despite their fictive origins, these legal ideas become real in a twofold sense: they affect the lives and material conditions of myriad individuals, and they invisibilize their origins by becoming hegemonic, decontextualized "natural" laws.

LLMs—as the algorithms powering ChatGPT, Bard, etc.—are headed in a similar direction, by attempting to create a universal view of the world based on (inherently) incomplete information. By presenting a monolithic, perspectiveless interface, chatbots and other surfaces for these models appear as a disembodied "natural" voice, obscuring their specific origins in a discrete set of training data. Even providers of these models caveat their use with warnings regarding the models' behavior, their outputs become "real" (both for future consumption by humans and, crucially, other models) in the sense that they are indistinguishable from material generated by humans. This creates feedback loops where a model's "opinion" can become a future reality.²⁴

A. Indigenous Data Sovereignty

The United States government and its agents have a long history of manipulating, misrepresenting, and outright falsifying information concerning Native Nations.²⁵ A number of treaties were ratified under dubious pretenses, with Native Nations "alleging that federal officials undercounted the total population and sometimes miscounted by including persons who were not even tribal members."²⁶

The power to define who you are is inextricably linked to data, beginning with the assertion that you exist at all.²⁷ Even today, programs like the census disproportionately

²³ Native Governance Center, *Blood Quantum and Sovereignty: A Guide*, (Feb. 25, 2024)

<https://nativegov.org/resources/blood-quantum-and-sovereignty-a-guide/>. "Blood quantum did not play a role in determining Tribal citizenship until the Indian Reorganization Act was passed in 1934. Under this federal law, many Native nations adopted boilerplate constitutions developed by the federal government that included using blood quantum as a basis for citizenship."

²⁴ See Christian, *infra* note 86, at 51 ("uncareful deployment of these models might produce a feedback loop from which recovery becomes ever more difficult or requires even greater interventions.") See also Emily M. Bender et al, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, ACM DIGITAL LIBRARY, <https://dl.acm.org/doi/10.1145/3442188.3445922>.

²⁵ See *Johnson v. M'Intosh*, 21 U.S. 543 (1823). From the very beginning of colonial contact, settlers and their eventual federal government chose to portray land as "unused" (or "unimproved," in a Lockean sense) and built legal structures to reify this interpretation. Widely regarded as fraudulent by members of the Cherokee Nation, the Treaty of New Echota is a particularly painful example of the United States misrepresenting its dealings with tribal nations. See also Dennis Zotigh, *The Treaty That Forced the Cherokee People from Their Homelands Goes on View*, SMITHSONIAN MAGAZINE (April 24, 2019), <https://www.smithsonianmag.com/blogs/national-museum-american-indian/2019/04/24/treaty-new-echota/>. "Another fraudulent treaty, the Treaty of Payne's Landing, in 1832, signed by a few nonrepresentative "chiefs," promised the Seminole land west of the Mississippi." TREUER, *supra* note 4, at 36.

²⁶ Tsosie, *supra* note 16.

²⁷ Particularly with respect to blood quantum, but also with respect to regulations around federal recognition, which require extensive documentation on the part of the Native Nation.

undercount Native Americans,²⁸ buttressing cultural myths (consciously or unconsciously) like the "vanishing Indian."²⁹ This question, of whether Indigenous peoples have "vanished," carries heavy legal import in modern times, with decisions like *McGirt*,³⁰ *Castro-Huerta*,³¹ and *Brackeen*,³² threatening to further erode rights and protections. This historical linkage between data and personhood carries into the present as a principal consideration of IDSov.

IDSov encompasses a set of frameworks (defined in different ways by different groups, from the Māori³³ to Native American coalitions³⁴) for placing control over data in the hands of Native Nations and Indigenous people. A catalyst in the development of this field has been the UN Declaration of Rights of Indigenous Peoples, which includes the right to "maintain, control, protect and develop their cultural heritage, traditional knowledge and traditional cultural expressions," alongside other cultural protections.³⁵

IDSov asks that we look at how data can construct group identities and hierarchies of power, pushing to recognize that data collection (whether it's a census, medical testing, or something else) is an inherently political activity, rather than a neutral one.³⁶ Frequently, dominant groups use the collection of data to justify and reinforce dominant power structures.

²⁸ Ben Kessler, *Native Americans, the census' most undercounted racial group, fight for an accurate 2020 tally*, NBC NEWS (Dec. 29, 2019) <https://www.nbcnews.com/news/us-news/native-americans-census-most-undercounted-racial-group-fight-accurate-2020-n1105096>. See also TREUER, *supra* note 4, at 380. In seizing land during allotment, "whites wanted the land and sent in a genetic investigator. In short order, the number of registered full-bloods at White Earth Reservation went from more than 5,000 to 408."

²⁹ Dina Gilio-Whitaker, *Real Indians, the Vanishing Native Myth, and the Blood Quantum Question*, ICT NEWS, (Sept. 12, 2018) <https://ictnews.org/archive/real-indians-the-vanishing-native-myth-and-the-blood-quantum-question>. Under the Dawes ("Allotment") Act of 1887, the federal government claimed "surplus" land on reservations after distributing fixed-size allotments to the "known" Native population. General Allotment Act of 1887, ch. 119, 24 Stat. 388 (codified as amended at 25 U.S.C. §§ 331–358). Thus, not only did undercounting harm Native Nations; but the federal government had a direct incentive to undercount. Treuer, *supra* note 4, at 83. The usual story told about us—or rather, about "the Indian"—is one of diminution and death, beginning in untrammelled freedom and communion with the earth and ending on reservations, which are seen as nothing more than basins of perpetual suffering."

³⁰ *McGirt v. Oklahoma*, 591 U.S. ___, (2020).

³¹ *Id.* at 2500; See also *Oklahoma v. Castro-Huerta*, 597 U.S. 629 (2022). In asking whether a large portion of the State of Oklahoma is, in fact, Indian territory, a throughline of the dissent in *McGirt* and the majority in *Castro-Huerta* is the idea that, roughly, demographics have changed and we must adjust the law to fit those demographics.

³² Opponents of ICWA and related policies benefit from the persistent undercounting and under-contacting of Native families as potential adoptive homes for Native children, persisting a fiction (rooted in white saviorism) where there are simply not enough Native families to adopt all the eligible children. *This Land: Solomon's Sword*, CROOKED MEDIA (Aug. 23, 2021) (downloaded using Google Podcasts).

³³ Te Mana Raraunga (Māori Data Sovereignty) Working Group, "TMR Māori Data Sovereignty Principles" (Oct. 2018) <https://static1.squarespace.com/static/58e9b10f9de4bb8d1fb5ebbc/t/5bda208b4ae237cd89ee16e9/1541021836126/TMR+Ma%CC%84ori+Data+Sovereignty+Principles+Oct+2018.pdf> (last visited Feb. 29, 2024).

³⁴ The Native Nations Institute at The University of Arizona has compiled numerous projects and recommendations regarding IDSov in the context of the United States. UNIV. OF ARIZ. NATIVE NATIONS INST.: <https://nni.arizona.edu/our-work/research-policy-analysis/indigenous-data-sovereignty-governance> (last visited Feb. 29, 2024).

³⁵ G.A. Res. 61/295, Art. 31 (Sept. 13, 2007).

³⁶ See KIM TALLBEAR, *NATIVE AMERICAN DNA: TRIBAL BELONGING AND THE FALSE PROMISE OF GENETIC SCIENCE*, University of Minnesota Press (2013) (explaining that the landscape of harms at the nexus of medicine and data is particularly fraught, including the potential for genetic testing to undermine tribal sovereignty).

Dr. Maggie Walter, Professor of Sociology at the University of Tasmania, describes this practice as gathering the "data of disregard,"³⁷ looking at the ways that selective data collection and representation can be used to pathologize communities (e.g. impoverished, low life expectancy, prone to substance abuse³⁸), and echoing of the idea of (and problems with) deficit framing³⁹ in other contexts.⁴⁰ By framing these narratives around the "objective" truth of data, dominant groups obscure the importance of power structures and historical context, perpetuating narratives that reinforce their hegemonic cultural norms.⁴¹

In looking at the overall relationship between Indigenous peoples and information systems, IDSov draws on ideas that have been developed throughout history, including both intra-tribal norms and challenges that have emerged after colonial contact.⁴² Though many of these ideas pre-date the rise of big data (including LLMs and other derived content), the last few decades have sharpened focus on the relevant harms and accelerated the need for Indigenous peoples to formalize these standards in a way that preserves their right to self-determination.⁴³ In particular, LLMs create both a diffusion of responsibility and a moral vacuum. When a model "speaks," it is unclear who is responsible for what it says. Such speech is devoid of moral consideration and replete with bias, at best mimicking patterns of interpersonal racism where the

³⁷ Maggie Walter, *The Politics of Data*, THE AUSTRALIA AND NEW ZEALAND SCHOOL OF GOVERNMENT, (2017), <https://legacy.anzsog.edu.au/documents/5326-maggie-walter-the-politics-of-data/file> (last visited May 20, 2023).

³⁸ See Chelsea Vowel, *The Stereotype of the Drunken Indian*, ÂPIHTAWIKOSISÂN (Oct. 10, 2012), <https://apihtawikosisan.com/2012/10/the-stereotype-of-the-drunken-indian/>.

³⁹ See Sam McVeety, *Decompiling Oppression #55*, DECOMPILING OPPRESSION GOOGLE GROUP, (Feb. 11, 2022), https://groups.google.com/g/decompiling-oppression/c/jLDCoCcLIq8/m/ZJA4if_sBAAJ (for a discussion of deficit framing and further resources.)

⁴⁰ See Rebecca Nagle, *The Story of Baby O—and the Case That Could Gut Native Sovereignty*, THE NATION, (Nov. 9, 2022) <https://www.thenation.com/article/society/icwa-supreme-court-libretti-custody-case/>, (last visited Mar. 2, 2024) <https://www.thenation.com/article/society/icwa-supreme-court-libretti-custody-case/> (returning to ICWA and Native children, another throughline of (white) plaintiffs' arguments has been the unfitness of Native parents to raise their children).

⁴¹ Reva Siegel, *Why Equal Protection No Longer Protects: The Evolving Forms of Status-Enforcing State Action*, 49 STAN. L. REV. 1111 (1996-1997). ("The essay demonstrates how efforts to dismantle an entrenched system of status regulation can produce changes in its constitutive rules and rhetoric, transforming the status regime without abolishing it."). By shifting the fulcrum of the conversation away from the actual harms perpetuated by settler colonialism and framing it instead as the individual and group failings of the communities harmed by those policies, white-led institutions can selectively provide aid (framed as altruism) in a way that does not fundamentally threaten their power.

⁴² Tsosie, *supra* note 16.

⁴³ See Stephanie R. Carroll et al., *The CARE Principles for Indigenous Data Governance*, 19 DATA SCI. J. 43, (2020), [DOI: 10.5334/dsj-2020-043](https://doi.org/10.5334/dsj-2020-043). (Example of recent scholarship on principles for Indigenous data governance.)

speaker has learned to not say the quiet part out loud.⁴⁴ Throughout, a rapid societal adoption curve⁴⁵ amplifies the impact of these concerns.⁴⁶

Beyond quantitative representations, qualitative information also can be a site of harm for Native Nations and their members. Post-contact, the United States waged a persistent war of misinformation regarding Native Nations, including the erasure of Native written language and agricultural practices in service of a narrative of uncivilized "savages."⁴⁷ In addition to misrepresentation, the undesired disclosure of information can harm Native Nations, such as traditional ecological knowledge that is disclosed in Freedom of Information Act requests during federal agency review processes.⁴⁸ In the landmark case, *Havasupai Tribe v. Arizona Bd. of Regents*, the Havasupai Tribe successfully sued over the use of blood samples for reasons beyond the original collection on the grounds that it violated their cultural property and creation stories.⁴⁹ Here, uncontrolled use of information had the potential to create lasting cultural harm.

DNA testing and blood quantum merit special attention here. The federal government has successively (1) defined Indigeneity in terms of blood quantum, (2) coerced Native Nations to use their definition in their IRA constitutions, and (3) called ancestry-based protections increasingly suspect.⁵⁰ Essentially, the government has forced its own definition of tribal membership upon Native Nations and then proceeded to progressively undermine the legal basis for protections based on that definition. Importantly, blood quantum itself is based on egregiously indeterminate (and racist) bases; historical determinations of who was a "full Indian" were based on racist scientific practices like phrenology.⁵¹ Again, the feedback loop between law

⁴⁴ See Sam Biddle, *The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques*, THE INTERCEPT, (Dec. 8, 2022), <https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/>. In a number of striking examples, early versions of ChatGPT declared the relative worth of different racial and gender identities. Although subsequent releases have added post-filtering to prevent these specific examples, the underlying model biases persist.

⁴⁵ See Geoffrey A. Moore, *Crossing the Chasm: Marketing and Selling High-Tech Goods to Mainstream Customers* (New York: HarperBusiness 1991). New technologies often display a growth trajectory that, when graphed, has a characteristic shape, beginning with gradual initial adoption and then sharp growth.

⁴⁶ ChatGPT is the "fastest-growing consumer application in history," having reached 100 million monthly active users in January 2023. Krystal Hu, ChatGPT sets record for fastest-growing user base — analyst note, Reuters, (Feb. 2, 2023), available at <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/> (last visited Mar. 2, 2024).

⁴⁷ See TREUER, *supra* note 4, at 30-70.

⁴⁸ See Sophia E. Amberson, Comment, *Traditional Ecological Disclosure: How the Freedom of Information Act Frustrates Tribal Natural Resource Consultation with Federal Agencies*, 92 WASH. L. REV. 937 (2017).

⁴⁹ Celeste Headlee, *Blood Victory in Medical Research Dispute*, NPR, (April 25, 2014), <https://www.npr.org/2014/04/25/306832661/blood-victory-in-medical-research-dispute> ("Researchers looked into mental illness, inbreeding, even migratory patterns that contradicted Havasupai traditional belief."). See also *Havasupai Tribe v. Arizona Bd. of Regents*, 220 Ariz. 214 (Ariz. Ct. App. 2009).

⁵⁰ In *Rice v. Cayetano*, the majority held that membership requirements designed to enfranchise Native Hawaiians ran afoul of the Fifteenth Amendment. At least one justice from the *Brackeen* majority has signaled a willingness to hear a challenge to the law under Equal Protection grounds (currently at a historic anti-classification low point, post-*Students for Fair Admission*).

⁵¹ TREUER, *supra* note 4, at 150. "For the purposes of determining who had European blood, teams of phrenologists were sent to Indian country. Using the theory and practice of craniometry developed by Samuel George Morton

and data obscures these origins (while retaining the outcomes), and documents such as census records are presented as fact, with their phrenological origins being lost to the "bit rot"⁵² of time. With IDSov highlighting the primacy of data for tribal sovereignty, the emergence of LLMs offers new dangers for Native Nations.

B. Large Language Models

Machine learning⁵³ is a decades-old discipline of computer science, focused on the goal of using automated systems (computers) to understand and reproduce patterns in ways that can be harnessed for augmenting human labor.⁵⁴ Rather than program explicit instructions into the machine ("if X, then do Y"), machine learning endeavors to have computers "learn" these patterns through extensive trial and error, ideally in ways that are not limited by the immediate imagination and cognitive bandwidth of the programmer. In contrast to the scientific method, which begins with a hypothesis and experimental design, the output of a machine learning process might suggest many distinct hypotheses for a given domain. An undiscerning observer may regard any of these hypotheses with unearned validity.⁵⁵

Herein lies the significant statistical danger. It is entirely possible to find meaningless correlations in data, and, lacking agency, a machine learning process will uncritically report all such correlations with equal weight.⁵⁶ Subsequently, a hasty—or ill-intentioned, or both—researcher can cherry-pick these findings (e.g. "people with blond hair have a drinking problem") to support the societal agenda of their choice.⁵⁷

Models are a core concept of machine learning, essentially acting as an oracle that provides answers ("output") when given questions ("input"). Models are created when a series of examples are fed through a computer algorithm (essentially a series of algebraic equations) to determine ("train") what the optimal configuration for that algorithm is "learning" the best

(which determined "scientifically" that Caucasians had the largest brain capacity—1,426 cubic centimeters versus a paltry 1,344 for American Indians and 1,278 for blacks), the phrenologists set about measuring skulls, and on this basis thousands of Indians were, surprisingly, deemed competent to own land: the better to steal it from them."

⁵² Bit rot refers to the "tendency for digital information to degrade or become unusable over time." *Bit rot*, MERRIAM-WEBSTER DICTIONARY, available at <https://www.merriam-webster.com/dictionary/bit%20rot> (last visited Mar. 2, 2024).

⁵³ Often intermingled with the term "artificial intelligence" (or "AI," and the futuristic promises it implies), some researchers used the term "machine learning" to scale back expectations towards a more incremental rate of progress. Rapid improvements in machine learning have brought the term AI back to the fore.

⁵⁴ As machine learning processes increasingly replace human labor, this arguably undermines the Lockean philosophical foundation of property as improved by human labor. Regrettably, I will only mention this topic in passing.

⁵⁵ For general background on how models surface results, see Christian, *infra* note 86.

⁵⁶ To date, LLM-based chatbots present their answers authoritatively, without a relative notion of confidence or uncertainty.

⁵⁷ See Luke Stark and Jevan Hutson, *Physiognomic Artificial Intelligence*, 32 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 922 (2022). See also Christian, *infra* note 86. Purveyors of facial recognition software have used correlations between appearance and identity to digitally revive the discredited ideas of physiognomy and phrenology.

approach in the process. For example, a cat recognition algorithm might be provided with a set of labeled images ("cat" and "not cat") and would learn a set of coefficients that allowed it to best predict whether an image contained a cat when a new picture is provided.⁵⁸ Some models can behave in generative ways as well, for example, in completing the analogy "colonialism is to empire as pluralism is to ???" by providing the missing word.⁵⁹

LLMs represent a sea change in the capabilities of machine learning models, primarily due to their flexibility and facility with language. These models are called "large" because their creators train them on much larger datasets than previous models. Their flexibility is illustrated by the presence of "foundation models," which are raw outputs of this initial training step that have shown themselves to be adaptable in multi-modal contexts (e.g. speech recognition, image comprehension). Previous models would tend to be specialized (as with the cat example, above), where a model would have little to no ability to perform well on a domain that it wasn't trained for. LLMs, however, can perform well on vastly different domains without starting over with a new training process.⁶⁰

Because these models are so adept with human languages, they blur the line between *learning* and *understanding*, often creating a false veneer of intelligence for human observers. Between the sincerity of LLMs,⁶¹ their facility with language, and automation bias,⁶² LLMs are custom built to "hack" human credibility biases⁶³ and increase the likelihood that consumers will accept uncritically whatever information the model provides. At their core, though, these models are simply the product of their inputs. When trained on a large corpus of text from the internet, models will absorb any biases present in this text and faithfully reproduce⁶⁴ them in model

⁵⁸ Jia Deng, et al., *ImageNet: A large-scale hierarchical image database*, INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, 248-255 (2009) <https://ieeexplore.ieee.org/document/5206848>.

⁵⁹ See Christian, *infra* note 86, at 40. In the author's testing, the answer was "diversity." Notably, these kinds of analogy tests are excellent at illustrating the bias in models. Uncensored models will readily complete things like "man is to doctor as woman is to nurse."

⁶⁰ See Ashish Vaswani et al., *Attention is all you need*, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (NIPS Proceedings 2017), (I. Guyon et al eds) (2017) ("We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.").

⁶¹ As noted previously, most models will present hypotheses with the utmost sincerity, leaving the operator to divine what is fact and what is fiction (or "hallucination," in the parlance of LLMs).

⁶² See Kate Goddard et al., *Automation bias: a systematic review of frequency, effect mediators, and mitigators*. 19 J. AM. MED. INFO. ASS'N 121 (2012) <https://academic.oup.com/jamia/article/19/1/121/732254> (last visited Mar. 2, 2024) ("The tendency to over-rely on automation.")

⁶³ This is exacerbated by human raters prioritizing the most plausible (not necessarily the most correct) answer during model training. See Ouyang *infra* note 66. See also Mary Cummings, *Automation Bias in Intelligent Time Critical Decision Support Systems*, Amer. Inst. Of Aeronautics and Astronautics (AIAA) 1st Intelligent Systems Technical Conference, AEROSPACE RESEARCH CENTRAL (Sept. 2004), <https://arc.aiaa.org/doi/10.2514/6.2004-6313>. ("Automation bias occurs in decision-making because humans have a tendency to disregard or not search for contradictory information in light of a computer-generated solution that is accepted as correct and can be exacerbated in time critical domains. Automated decision aids are designed to reduce human error but actually can cause new errors in the operation of a system if not designed with human cognitive limitations in mind.")

⁶⁴ See Sam Biddle, *The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques*, THE INTERCEPT (Dec. 8, 2022), <https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/>.

outputs.⁶⁵ Recent models have added human feedback as another major component in their training input.

A key advancement in the perceived intelligibility of these models is a technique known as "reinforcement learning with human feedback" (RLHF), where human raters grade the performance of a model; these signals are then used to re-train the model to bias toward behaviors that would be rated highly.⁶⁶ While this process attempts to incorporate some objective measures (e.g. "is this answer factual, given the following references?"), it incorporates a further layer of subjectivity into the model, as raters impose their own conceptions and world views into their ratings (particularly with highly qualitative metrics like "readability" and "tone" of an answer). One reason for these techniques is the difficulty of creating robust benchmarks for models; there are numerous problems with artificial benchmarks being uncorrelated with human evaluation and/or favoring responses of a particular type of model (with a clear "real" difference in quality).

LLMs have already inspired numerous legal challenges related to copyright and authorship. For example, many contributors to the online code repository Github believe that their work has been unfairly used to benefit the creation of some foundation models (an assertion that is reinforced by the phenomenon of "recitation," where a model reproduces, verbatim, one of its inputs).⁶⁷ The picture is blurrier for inexact copies (e.g. a picture in the style of a given artist). A clear legal precedent and well-defined framework for distinguishing illegal copying and fair use do not yet exist in this context. Given the centrality of information and data to IDSov, the outcomes of these cases will have important consequences for Native Nations.

There are clear implications for Indigenous economic and cultural concerns, including the potential "digital appropriation" of Native art and traditions to the perpetuation of structural harms against Native Nations.⁶⁸ By training a widely deployed model on a large corpus of art, a model becomes capable of credibly responding to prompts like "show me an eagle in the style of the Duwamish" without honoring (or even being aware of) the corresponding artistic norms, cultural practices, and beliefs of the Native Nation. In addition to the cultural insult, scenarios like these also directly dilute tribal economic power and agency to choose what art to create and how and whether to sell it. Thus, LLMs have sweeping implications for the behavior of data in general, and specifically for Native Nations. In the next section, I explore these connections in more depth.

⁶⁵ See Birhane, *supra* note 16. As noted earlier, the situation may, in fact, be worse, with models amplifying biases in their training set.

⁶⁶ See Long Ouyang et al., *Training language models to follow instructions with human feedback*, ARXIV (2022), <https://arxiv.org/abs/2203.02155>.

⁶⁷ See Saveri, J. R., & Zirpoli, C. Class action complaint. United States District Court for the Northern District of California, https://githubcopilotlitigation.com/pdf/06823/1-0-github_complaint.pdf (filed November 3, 2022). See also *Doe 1 v. GitHub, Inc.*, No. 22-CV-06823-JST, 2023 WL 3449131, at *1 (N.D. Cal. May 11, 2023) (for details on the complaint).

⁶⁸ Indian Arts and Craft Act (IACA), 25 U.S.C. § 305-305(e) (1990). The current version of IACA would be insufficient for this task, given that it only lightly protects authorship claims through regulating marketing.

C. Linking the Law, IDSov, and LLMs

There are parallels between how the United States legal system operates and the training and feedback loops driving LLMs. Returning to the example of blood quantum, I have shown how the collaboration of the legal and data fields can reify a concept as "objective" when it has no such factual basis. Similarly, the training and refinement process for LLMs depends on a subjective rating system that reflects the biases of its users. However, when compared against other models via benchmarks, that process becomes "objective" as a snapshot of the model's performance.⁶⁹ Consider the evaluation question "how accurately does this response represent Indian culture?" This can become a virtual measure akin to blood quantum, with models rated on relative ability to make a determination of "Indianness". Both are artificial evaluations, with the former created to reinforce colonial power structures, and the latter created by profit motives to have an "authentic" model. The underlying artificiality of the concept, and the absence of any "real" data to ground on, disappear in the feedback loop; one model is "better" than another, even if there is nothing to actually be "better" at.

Clearly, a body of law is needed to govern the behavior of LLMs. To inform these governing principles, previous analyses of Federal Indian law provide a useful set of parallels. In particular, Professor Maggie Blackhawk⁷⁰ suggests that Federal Indian law become a dominant paradigm for examining public law (in particular, civil rights law) in the United States at large. She contrasts Federal Indian law with the dominant paradigm of protecting minoritized populations from majority rule with rights frameworks and protections against majoritarian abuse of power.⁷¹ These rights frameworks typically fail to contemplate the actual cession of power to minoritized populations; power only interacts with these populations as something to be protected *from*. Conversely, the paradigm of Federal Indian law and tribal sovereignty features such cessions of power as central to its operation and ability to protect tribal interests.⁷² In the realm of LLMs, prevailing notions of minority protection hew towards the former paradigm, with guardrails (typically weak, at that) serving to protect minoritized interests *from* potential harms, without granting any actual agency over the model operation. A better governance model for LLMs would incorporate tribal sovereignty as a core principle.

Another important parallel from Professor Blackhawk's discussion of paradigms contemplates the means for prevention and redress of harms. The rights-based paradigm assumes that minoritized populations must assimilate into majority legal and governance structures in

⁶⁹ For one such benchmark, *see* Christian, *infra* note 86, at 31.

⁷⁰ Fond du Lac Band of Lake Superior Ojibwe, NEW YORK UNIVERSITY SCHOOL OF LAW, <https://its.law.nyu.edu/facultyprofiles/index.cfm?fuseaction=profile.overview&personid=56076> (last visited April 7, 2024).

⁷¹ Maggie Blackhawk, *Federal Indian Law as Paradigm Within Public Law*, 132 HARVARD L. REV. 1787, (2019) <https://ssrn.com/abstract=3486307> (last visited Mar. 2, 2024).

⁷² *Id.* at 1790. "Unlike slavery and Jim Crow segregation, federal Indian law teaches that nationalism is no panacea for majority tyranny, and that rights can wound as well as shield minorities."

order to avail themselves of those protective rights.⁷³ For LLMs, the dominant supposition of the day is that regulation of such models will be accomplished via government (rights-based) regulation and corporate advocacy, often with the corporation setting the rules for such engagement. This artificially forecloses the possibility of discussing and preventing harms that are not contemplated by existing corporate or government-defined structures.

In viewing IDSov through the paradigm of Federal Indian law, the history of allotment offers a key insight on the nexus of data and law.⁷⁴ Although motivated by a tremendous amount of greed, it is important to remember that many white advocates of allotment also framed themselves as "saviors of the Indians" (drawing heavily upon the data of disregard in making that case). This same type of saviorism (producing well-meaning solutions that do not actually center the desires of the people affected) are apt to reappear in LLM governance. Further, existing feedback models for LLMs are highly individualized in nature, erasing the possibility of group harm from the dialogue, and perpetuating the same kind of individualized frame as allotment.⁷⁵ Despite the severity of these harms, the federal government often denies Native Nations adequate redress.

III. HARMS TO TRIBAL INTERESTS

The United States has a long history of distorting and minimizing harms to Native Nations and their respective interests. Even when some form of harm is recognized, Native Nations must render it legible to a legal system rooted in coloniality.⁷⁶ Native Nations have pushed back on these narrow conceptions of harm and restitution, including the Lakota People's demand for the return of the Black Hills.⁷⁷ LLMs are likely to trigger the same issues, superficially only implicating considerations of harmful speech and representation. Current jurisprudence categorically denies the significance of these harms.⁷⁸ In the following section, I discuss the ways that (1) data practices can amplify existing disparities, (2) lack of data can perpetuate harm, and (3) LLMs can create culturally specific harms.

⁷³ In a colonial context, this assimilation only furthers the colonial project and accelerates Indigenous erasure. "Forcing colonized peoples to engage in our democratic process to avoid subordination only furthers the colonial project." *Id.* at 1800.

⁷⁴ See Dawes Severalty Act, *supra* note 7.

⁷⁵ See Blackhawk, *supra* note 71 at 1853. More pointedly: "Rights were harbingers of neoliberal individualism and destroyers of the very communities who ensured our humanity."

⁷⁶ As with the Indian Claims Commission, which narrowly interpreted many complaints and typically offered minimal monetary restitution. See Treuer, *supra* note 4; Indian Claims Act of 1946, Pub. L. No. 79-726, 60 Stat. 1049 (1970) (codified at 25 U.S.C. §§ 70a-70v).

⁷⁷ LAKOTA PEOPLE'S LAW PROJECT, <https://action.lakotalaw.org/action/land-back> (last visited Mar. 19, 2024).

⁷⁸ For a discussion of the harms of subordinating speech, see Mari J. Matsuda, *Public Response to Racist Speech: Considering the Victim's Story*, 87 MICH. L. REV. ASS'N. 2320 (1989), <https://doi.org/10.2307/1289306>.

A. The Data of Disregard Amplifies Existing Disparities

Colonial data gathering practices exhibit a "data of disregard" pattern, where facially objective data projects can perpetuate stereotypes and a deficit frame.⁷⁹ When combined with the data ingesting behavior of LLMs (that is, that they reproduce the biases of their input data), it is almost inevitable for the data of disregard to lead to models that reproduce the societal biases of their creators. Because the training process for the model has no robust notion of credibility or context to draw on (it is all "just data"), all training data provided to a model is equally and credulously regarded as fact. There is simply no "reality check" mechanism to flag and problematize suspect findings in a model.⁸⁰ Thus, the "data of disregard" that characterizes many data projects around Indigenous communities today would likely result in model biases in economic,⁸¹ medical,⁸² and other fields.

In examining the harms of the "data of disregard," child welfare is one area where these effects could be particularly harmful. In Pittsburgh, a child welfare agency computed automated risk scores when considering child placements, resulting in an opaque system with limited options for legal recourse.⁸³ In this case, the information feeding into the model was at least known to the agencies, but for currently available foundation models, the data sources are entirely discretionary.⁸⁴ Assuming that at least some of the "data of disregard" has made it into these models, the outcomes will almost certainly reflect these societal biases.⁸⁵ While the training and operation of such models typically forbids *explicitly* predicating risk factor outcomes on protected characteristics like race, they can reach nearly identical conclusions by depending instead on *proxy* variables (e.g. zip code) that closely correlate with protected characteristics.⁸⁶

Unfortunately, this reproduces, in digital form, what has already been happening societally for centuries with neglectful (and often purposely inaccurate) information. Under the status quo, LLMs will already reproduce the elisions and erasures already present in the data currently collected by society. As such, the underinvestment in data that is currently collected will be the default representation available in these models. Further, for those marginalized even

⁷⁹ See Walter, *supra* note 37.

⁸⁰ The practice of "grounding" a model, where the model fact-checks against a corpus of data, typically only affects direct factual citations and cannot handle abstract inferences based on multiple pieces of information. For more information on grounding, see Serge Thill et al., *On the Importance of a Rich Embodiment in the Grounding of Concepts: Perspectives From Embodied Cognitive Science and Computational Linguistics*, 6 TOPICS IN COGNITIVE SCIENCE 545 (2014).

⁸¹ While classification-based discrimination in credit checks, for example, is illegal, the proliferation of algorithms to power the gig economy and other extra-regulatory domains would be difficult to detect without extensive resources for research, and even harder to legally challenge.

⁸² As mentioned earlier, everything from cultural practices to medical treatments to sovereignty claims.

⁸³ See EUBANKS, *supra* note 15, at 127-168.

⁸⁴ See *Id.*

⁸⁵ Were it favorable to their cause, future challengers of ICWA would almost certainly attempt to marshal and decontextualize tools like this in an attempt to portray Native families as unfit and dangerous for children.

⁸⁶ For a deeper discussion of proxy variables in machine learning, see BRIAN CHRISTIAN, *THE ALIGNMENT PROBLEM: MACHINE LEARNING AND HUMAN VALUES* (W.W. Norton & Company, 2020).

within Indigenous groups (e.g. non-affiliated individuals and non-recognized tribes) these patterns would be reproduced and amplified.

B. The Double-bind of Erasure vs. Misrepresentation

Even some of the potential mitigations for digital harm carry risks with them. While I consider solutions more thoroughly in a subsequent section, here I look at approaches that may actually make the problem worse. Given the current LLM landscape, perhaps the most obvious countermeasure is to attempt to label all Indigenous data as such and remove it from training, avoiding the uncontrolled reproduction and manipulation of this data.⁸⁷ Assuming that this could be done comprehensively, it would prevent the kind of cultural and economic appropriation discussed earlier. However, this would constitute a kind of digital erasure that has its own set of downsides, some obvious and others perhaps to be discovered.

For one, the absence of Indigenous cultural markers from the "knowledge" of an LLM constitutes a digital reproduction of the myth of the vanishing Indian. The real-world consequences of this would likely span many domains, from the cultural (absence of Indigenous representation in digitally created artifacts: "show me a family"⁸⁸) to medical (medical models' inability to reckon with conditions correlated with specific tribal subpopulations). Importantly, the economic consequences are particularly significant within the medical field, since access to medical care is limited within reservations, where this would further inflate costs and affect tribal health care systems. Culturally, this could create a negative feedback loop around the already lacking (frequently harmful) media portrayals of Native Americans, which often either reproduce (through omission) the "vanishing Indian" or perpetuate active stereotypes.⁸⁹

Broadly, the absence of Indigenous cultural markers creates a double bind around Indigenous attempts to manage the panopticon of LLM training.⁹⁰ Within existing structures,

⁸⁷ Current models are fed vast quantities of data. See Vaswani, *supra* note 60.

⁸⁸ In the author's testing, ChatGPT described the "typical" Smith family (John is an engineer; Sarah is a school teacher), with two kids and a dog. Notably, Google's Bard provided a much more inclusive response, including entries on same-sex, blended, multigenerational, and single-parent families. Attempts to force a standard model of family structure into child welfare cases has had profoundly harmful consequences for tribal members and their families, see Indian Child Welfare Act (ICWA), 25 U.S.C. §§ 1901-1963 (1978).

⁸⁹ With the proliferation of LLM-generated and -assisted content in the media, existing erasures and stereotypes will be reproduced ad infinitum. See Juana Majel Dixon, *Misappropriation of Native Identity in Film & Television*, NATIONAL CONGRESS OF AMERICAN INDIANS (2017), <https://www.ncai.org/resources/resolutions/misappropriation-of-native-identity-in-film-television>; Crystal Echo Hawk, *Indigenous Representation is Still Scarce in Hollywood: 'We Need More Native Stories'*, VARIETY (Oct. 11, 2021), <https://variety.com/2021/film/opinion/indigenous-representation-hollywood-native-stories-1235086445/> (last visited Mar. 19, 2024).

⁹⁰ As a specific illustration of this, consider LLM interactions with tribal codes. As of this writing, ChatGPT professed to be unable to summarize either the Navajo Nation or the Yakama Nation's code, using the same justification ("Although I do not have access to the specific details of the Navajo Tribal Code due to my knowledge cutoff date in September 2021..."). In fact, the Navajo Nation's code *is* online (<https://www.nnols.org/navajo-nation-code/> as of 2010), while the Yakama Nation states, "The Nation has not given permission to put the full text online," (<https://narf.org/nill/codes/confederatedyakama/index.html>). Conversely, both Bard and ChatGPT offer a summary

both participation and non-participation in training wrest away Native Nations' ability to self-define, threatening stereotypes and appropriation on one hand, and erasure on the other. This double bind runs parallel to many other challenges in Indigenous economic development.⁹¹ While Native Nations often face economic imperatives to interact with the world outside of the reservation (e.g. contracting with companies for resource extraction; attracting patrons to tribal casinos; creating art that is attractive to the white gaze), they risk cultural and societal compromise by trying to create institutions and attractions that are legible to non-Indigenous observers.⁹² At the same time, if Native Nations decide to ignore these external stakeholders, they are left with reservations that are often devoid of natural resources and with little recourse other than to engage with non-Indigenous entities to promote the health and well-being of their communities.

C. LLMs and Harm to Tribal Cultural Interests

Given both the apparent universality and the elimination of context that LLMs feature, there are numerous possibilities for cultural harms.⁹³ If one were attempting to learn about Native culture (whether Native or non-Native) from an LLM, it is likely that key contextual markers would be removed from traditional stories and practices.⁹⁴ Even ostensibly "close reproductions" of these artifacts may contain crucial errors and omissions.⁹⁵ Importantly, tribal law recognizes cultural harms that are not acknowledged by federal law. As a result, federal law is not equipped to safeguard tribal interests in LLMs.⁹⁶

of the US Code when prompted. Assuming that this type of access is useful for legal scholars, the double-bind lies between "put the code online for everyone so we can use an LLM" and "forgo the utility of this tool."

⁹¹ Others have noted the prevalence of double-binds in Indian law beyond economic development and representation. *See also* Eric Cheyfitz, *The Colonial Double Bind: Sovereignty and Civil Rights in Indian Country*, 5 U. PA. J. CONST. L. 223 (2003).

⁹² For an illustration of this tension and the frustrating tendency of outside observers to dictate "best practices" for Indigenous economic development, *see* Tamera Begay & Matthew L.M. Fletcher, *Ma'ii and Nanaboozhoo Fistfight in Heaven*, SO. CAL. L. REV. (forthcoming), U. MICH. PUB. LAW RESEARCH PAPER No. 22-050 (2022).

⁹³ "Lenore Keeshig-Tobias, a writer from one of Canada's First Nations, defines cultural appropriation as a "taking, from a culture that is not one's own, intellectual property, cultural expressions and artifacts, history and ways of knowledge." Tsosie, *supra* note 16, at 300.

⁹⁴ "In fact, in some tribal cultures, storytelling is reserved to certain seasons, and there may be controls on who is considered eligible to tell a specific story." Tsosie, *supra* note 16, at 303.

⁹⁵ "There may be constraints as to the season during which a ceremony may be performed, for example, and many Native Nations believe that even one mistake in the wording of a song or the placement of a design during a ritual can have profound and negative consequences. Many Indian people therefore find it completely unacceptable when unauthorized persons, who are not even familiar with the Native Nation's culture and language, pretend to "practice" their religions for either "spiritual" or "commercial" benefit." Tsosie, *supra* note 16, at 314.

⁹⁶ This includes defamation of the dead. "For example, the defamation claim would fail under Anglo-American law because of the cultural belief that the dead cannot be harmed. On the other hand, within the Lakota belief system, the spirits of the dead are very much alive, and they are perceived to have feelings and volition. Thus, the plaintiffs made a claim for "defamation of the spirit" under tribal customary law." Tsosie, *supra* note 16, at 352.

In examining the possible cultural harms listed by Professor Rebecca Tsosie⁹⁷ in her work on IDSov, all of them coincide with possible LLM behaviors. Briefly, the harms are (1) removal of cultural/sacred objects, (2) appropriation of cultural/sacred symbols for commercial use, (3) incorporation of cultural/sacred symbols into "New Age" religions, and (4) portrayal of Indian people in media created by non-Indians.⁹⁸ The very act of training a model arguably constitutes a violation of the first harm, as symbols are removed from their original context (and any corresponding restrictions on their usage). Instances of the second and third harm abound in third-party use of models, with non-Indians free to ask for "Indian-like" symbols, rituals, and artifacts, with the presumptive defense that these were inventions of the model, not "real" appropriation.⁹⁹

As far as storytelling and portrayals of Indians by non-Indians, LLMs offer nearly limitless (disturbing) possibilities. Touted for their ability to generate text and even do basic "creative writing" tasks, it is easy to imagine LLMs being used for a cheaply produced explosion of ostensibly Indian-centric stories. Beyond the harms of denying Native storytellers the commercial opportunities to tell these stories, the cultural impact of this kind of proliferation (animated by the underlying data of disregard) is hard to overstate. For non-Indians (including Supreme Court justices like Justice Rehnquist¹⁰⁰), such stories stand to reinforce a deficit-based, paternalistic view of Native Nations. For Indian consumers, it reinforces internalized oppression and intergenerational harm. For Native Nations seeking federal recognition, presentations from an LLM that tend toward an assimilationist frame could undermine those claims.¹⁰¹

Language merits special consideration here, given its centrality to both Indian cultures and LLMs. While LLMs are most well-known for their facility with written text, many are capable of speech synthesis (and transcription) as well. In the same way that allotment-era reformers sought to "save" the Indian through the actions that ultimately proved tremendously harmful, contemporary actors may seek to employ technology towards the preservation of native languages in a paternalistic, coercive way, experimenting with LLM behavior to generate and infer passages in Native languages. Not only does this create harm through the elision of context and subtle errors, but it also devalues and decenters the urgent, difficult work that is happening in many communities to preserve languages on their own terms.

Existing legal structures in the United States do not recognize the types of cultural harms detailed by Professor Tsosie. Further, attempts to shoehorn such cultural harms into existing

⁹⁷ "Rebecca Tsosie is a Regents Professor at the James E. Rogers College of Law at the University of Arizona ... Professor Tsosie serves as a Supreme Court Justice for the Fort McDowell Yavapai Nation and as an Associate Judge on the San Carlos Tribal Court of Appeals." THE UNIV. OF ARIZ., <https://law.arizona.edu/person/rebecca-tsosie> (last visited Mar. 19, 2024).

⁹⁸ *See id.*

⁹⁹ Not that real appropriation has fared much better, legally. Attempts to bar the use of "Crazy Horse" as a brand name have been struck down on First Amendment grounds, despite affirmative Congressional action. *See Hornell Brewing Co., Inc. v. Brady*, 819 F. Supp. 1227 (E.D.N.Y. 1993).

¹⁰⁰ *See* earlier comment on his dissent, Tsosie, *supra* note 16.

¹⁰¹ "For the Mashpee, of course, the court's finding that they were not an "Indian tribe" was crucial, because it meant that the protections of federal Indian law for tribal lands and cultures were unavailable to them. So, once again, the non-Native media "image" of Indians caused a tangible harm to the descendants of a historical Native group." *Id.*

modalities like property law only perpetuate the colonial project of Native assimilation and erasure, undermining sovereignty in the process.¹⁰² Anticipating speech-based defenses of this kind of appropriation, Professor Tsosie discusses various formulations of the identity-less "cosmopolitan" citizen, who freely samples from innumerable cultures in constructing their lifestyle. In many ways, the current positioning of LLMs frames them as the ultimate "cosmopolitan" amalgam of information, offering authority on all cultures while belonging to none and implicitly reflecting majority viewpoints.

IV. LEGAL FRAMEWORKS GOVERNING IDSOV AND LLMs

Outside of tribal courts, the legal landscape for defending tribal sovereignty can be daunting. The Court frequently defers to their own (selectively informed) perception of the actual state of things and a desire to avoid confusion for a nonmember-majority populace.¹⁰³ In the following section, I will examine the shortcomings of individual rights frameworks, the possibilities of tribal legal remedies, and other sources of data governance policies.¹⁰⁴

A. Yakima, Checkerboarding, and Individual vs. Collective Rights

During the allotment era, the federal government adopted a "divide and conquer" approach to Native Americans. The government divided reservations into individual parcels ("allotments") and distributed them to individuals, fracturing the Native Nation's overall control of the area.¹⁰⁵ Similarly, the distinction between individual and collective tribal property looms large in the fight against the use of tribal cultural property in LLMs. These algorithms stand to provide a dangerous stepping-stone towards erasing cultural property, decontextualizing information and culture in a way that renders the line between individual and collective property intractable within the current legal system.

As Native Nations and Native individuals evaluate responses to the use of tribal cultural property in LLMs, a possible solution is to develop codes that emphasize collective direction and ownership for information and data, lest they find themselves in a "digital allotment" situation where courts determine that individually held cultural property has been given away, foreclosing

¹⁰² "Native people lose the right to exclude others from cultural knowledge or condition their use of such knowledge. Once made available to the public, cultural knowledge (including songs, symbols, and stories) is perceived as the "common" property of the dominant society and is incorporated into that society's cultural expression with a different set of meanings and forms." *Id.* at 313-314.

¹⁰³ See *McGirt v. Oklahoma*, 591 U.S. ___, 140 S. Ct. 2452 (2020).

¹⁰⁴ See *Cnty. of Yakima v. Confederated Tribes & Bands of the Yakima Nation*, 502 U.S. 251 (1992). See also arguments in *McGirt*, 591 U.S. ___ (2020), regarding the effects on people who would be purportedly shocked to find themselves "suddenly" in Indian Country.

¹⁰⁵ Over time, fractionation (division of allotment ownership between offspring) has diluted ownership in individual parcels to absurd levels, making them even more vulnerable to predatory purchases.

the Native Nation from recourse. Art is perhaps the most obvious example, though one can imagine other sectors of similar import, such as ecological and other knowledge.¹⁰⁶

An illustrative consequence of allotment occurred in *Cnty. of Yakima v. Confederated Tribes & Bands of the Yakima Nation*, where the Court considered the applicability of an ad valorem tax on reservation land.¹⁰⁷ Facially, this would be a violation of tribal sovereignty, but the Court ruled against the Native Nation, in part due to the diminished and "checkerboarded" status of the reservation.¹⁰⁸ This checkerboarding is the specific consequence of allotment (rooted itself in the misperception that it would "save" the Indians), where the legal fictions of blood quantum and competency determined land ownership, which here results in specific harm to tribal sovereignty.

If a court finds that the legal and intellectual property landscape for a Native Nation's claims is similarly complex and diffuse, similar logic could apply. This would reduce opportunities for economic development not only for the Native Nation as a whole, but also for affiliated individuals to engage in culturally privileged economic activity if such material and techniques are deemed to be essentially in the equivalent of the public domain for algorithmically generated outputs.

In general, this would match the general neoliberal trend towards individualizing rights and remedies, one that continues to threaten tribal sovereignty.¹⁰⁹ The more that anti-tribal interests can portray Native Nations as simple collections of culturally adjacent individuals, legally and culturally, the greater the threat to group-based protections. When confronted with such failings of legal tools, the Court all too frequently defers to entrenched systems of power.

B. Legal Remedies Against Non-Tribal Members

Ideally, Native Nations would have robust legal means at their disposal to preempt harm in the current technical landscape. However, the current legal landscape restricts tribal jurisdiction over nonmembers. As far as whether Native Nations can take large-scale legal action against OpenAI (the creators of ChatGPT) or another supplier of LLMs, *Montana v. United*

¹⁰⁶ For an extensive survey of protection of tribal cultural property in tribal codes (including 49 Native Nations with Indigenous Data Sovereignty provisions), see Angela Riley, *The Ascension of Indigenous Cultural Property Law*, 121 MICH. L. REV. 1 (2022).

¹⁰⁷ 502 U.S. 251 (1992). Note also how the colonial spelling of "Yakama" has become "real" in the historical record, despite the Nation's actions to change it. YAKAMA NATION, *Yakama Nation History*, <https://www.yakama.com/about/> (last visited Mar. 18, 2023).

¹⁰⁸ "We cannot resist observing, moreover, that the Tribe's and the United States' favored disposition also produces a "checkerboard," and one that is less readily administered..." *Yakima*, 502 U.S. at 265 (1992).

¹⁰⁹ For the latest contortion of this logic with respect to race-based admissions, see *Students for Fair Admissions v. President and Fellows of Harvard College*, 600 U.S. 181, 190 (2023). "A benefit to a student who overcame racial discrimination, for example, must be tied to *that student's* courage and determination. Or a benefit to a student whose heritage or culture motivated him or her to assume a leadership role or attain a particular goal must be tied to *that student's* unique ability to contribute to the university. In other words, the student must be treated based on his or her experiences as an individual—not on the basis of race." *Id* at 231. Although Native rights are ostensibly distinguished from race-based selection and protected under *Morton v. Mancari*, that protection may not last under the current Court. 417 U.S. 535 (1974).

States has emerged as the Court's favorite test in the current era for nonmembers in Indian Country.¹¹⁰ With *Montana*, the Court established only two routes for Native Nations to pull nonmembers into tribal court: (1) when the nonmembers enter a "consensual relationship" with the Native Nation or tribal members and (2) when "conduct threatens or has some direct effect on the political integrity, the economic security, or the health or welfare of the Native Nation."¹¹¹

In order to make a framework under *Montana* workable, it falls to Native Nations to create economic incentives for engagement and "consensual relationships," and likely the federal government to require that Indian Country not receive differential treatment with respect to nonmember services. An explicit agreement between nonmember corporations and Native Nations would be a best-case scenario.¹¹² While it is possible that Native Nations may be able to enter into agreements with OpenAI and other entities for allowed use of data, it is far more likely that the data will be used without the consent of the relevant Native Nation(s) or a clear means of legal recourse.

For non-consensual use of data and representation, the second exception ("conduct threatens or has some direct effect on the political integrity, the economic security, or the health and welfare of the tribe") might offer protection. Textually, this is a promising avenue; one can readily muster arguments of how misrepresentation (whether by erasure or pathologization) threatens nearly all of the above. Inaccurate and/or incomplete medical representations threaten the health of the Native Nation; acontextual economic studies and evaluations similarly threaten economic security; etc. However, the Court has given little reason so far to believe that this legal avenue will bear fruit, having only recognized valid instances of these exceptions in a handful of cases.¹¹³ For suits outside of Indian Country, legal options are currently limited, although the federal government could provide a cause of action (building on frameworks like the Native American Graves Protection and Repatriation Act (NAGPRA) and the Indian Arts and Crafts Act (IACA).

Other avenues are similarly limited. While data centers themselves are unlikely to lie on tribal land,¹¹⁴ internet access is present on many reservations (though often severely inadequate).¹¹⁵ Whether this meets the *International Shoe*¹¹⁶ standard for minimum contacts is

¹¹⁰ See *Montana v. United States*, 450 U.S. 544 (1981).

¹¹¹ *Id.* at 565.

¹¹² If nonmembers enter into consensual relationships with Native Nations, that can be a basis for extending jurisdiction, although this has been inconsistently applied. For example, in *Atkinson*, the Court held that a hotel occupancy tax on nonmembers was invalid. *Atkinson Trading Co. v. Shirley*, 532 U.S. 645, 121 S. Ct. 1825, 149 L. Ed. 2d 889 (2001).

¹¹³ Regarding the Court's resistance to characterizing speech as harmful, see Matsuda, *supra* note 78.

¹¹⁴ While not on a reservation, data centers at the Dalles site in Oregon occupy traditional Indigenous cultural sites. Jeremy Takala & Lauren Goldberg, *Stop sacrificing Indigenous sacred sites in the name of climate change*, THE SEATTLE TIMES (Sept. 25, 2022), <https://www.seattletimes.com/opinion/stop-sacrificing-indigenous-sacred-sites-in-the-name-of-climate-change/> (last visited Mar. 19, 2024).

¹¹⁵ *Exploring the Lack of Internet Access on Native American Reservations*, COMMUNITY TECH NETWORK (July 28, 2023), <https://communitytechnetwork.org/blog/exploring-the-lack-of-internet-access-on-native-american-reservations/> (last visited Mar. 19, 2024).

¹¹⁶ See *International Shoe Co. v. Washington*, 326 U.S. 310 (1945).

another story; Native Nations likely face an uphill battle to show that OpenAI and others have purposely availed themselves of particular tribal fora. Additionally, this raises a similar double-bind concern, given that companies could intentionally block their services in Indian Country to avoid any appearance of intentional engagement.

C. Data Protection in the European Union

The EU currently leads the world in comprehensive regulation of digital industries. While regulations explicitly targeting LLMs are not currently in effect, the EU has amended proposals for the upcoming AI Act to include a disclosure requirement for model creators, a concept that should be adopted broadly.¹¹⁷ Such a provision requires that the creators of models at least disclose the training data on which the model was trained, yielding numerous benefits. The United States could adopt similar disclosure provisions as a legal framework to protect tribal interests and demand transparency of data used to generate LLMs.

Most immediately, such a disclosure eliminates the "guessing game" of trying to determine whether potentially sensitive information was included in a model (whether copyrighted or tribal cultural property), which currently must be done through output-based testing, with no guarantee of accurate results. This kind of paper trail makes the path to legal action much clearer since it would indicate what *actually* is in a model, rather than requiring court challenges (with no guarantee of success) to determine the inclusion of a given artifact on an individual basis.

A disclosure provision also provides benefits towards adversarial testing around bias and representation. In adversarial testing, a third party can probe the behavior of a model using "what if?" constructions (for example, altering specific identity-related aspects of the input) to determine whether there is differential behavior for different identity groups. When the source data for a model is available, checks on model output can be tailored more specifically to the input of the model (for example, to see whether the inclusion of certain inputs creates biases in the outputs). This would be especially important in the case of data that is *not* tribally produced, but, nonetheless, has import related to Native Nations (e.g. census data), for example, if a Native Nation wished to assert that the inclusion of such data was creating biased outputs in violation of tribal code.

V. PROTECTING TRIBAL SOVEREIGNTY IN THE AGE OF LLMs

In developing solutions, it is important to avoid the mistakes of the past (and present). Policy recommendations should be developed in direct collaboration with tribal stakeholders at every step. Possible solutions include creating tribally controlled data and models, adopting protective data handling and training practices, and strong, informed federal action to enforce tribal

¹¹⁷ PRESS RELEASE, EUROPEAN PARLIAMENT, *AI Act a step closer to the first rules on Artificial Intelligence* (May 11, 2023), <https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence> (last visited Mar. 19, 2024).

interests. As solutions are evaluated, it is critical to preserve tribal sovereignty and cultural integrity.¹¹⁸

A. A Tribal Sovereignty Approach to LLMs

Tribal sovereignty should be at the heart of any approach to managing LLMs. Just as Native Nations can codify their own laws and convene courts, Native Nations should be in control of the information that defines them in the eyes of themselves and others. Drawing again on federal Indian law as an instructive paradigm, the current approach of having enormous, singular foundation models should give us pause because of its lack of contemplating cession of power across a plural society. Incorporating "rights" into such centralized models is likely insufficient for preventing results that further the subordination of Native peoples, and policymakers should instead contemplate what true cession of power would look like in this context. Most obviously, the creation of many complementary models, each informed by a different corpus of data and training techniques and *under the control* of respective Native Nations, would most closely hew to the current model of tribal sovereignty.

If Native efforts to rehabilitate digital representations are successful, they could pave the way for new models of data handling that accrue benefits beyond Native Nations and tribal members. This would parallel the ways that Native Nations are leading with "new" (actually often traditional, as in, since time immemorial) practices in everything from forest management¹¹⁹ to medical treatment,¹²⁰ carefully navigating the importance of preserving cultural integrity along the way. Basic protections around consent for sharing could adopt and improve upon EU requirements (current and proposed), bringing a much-needed concept of group identity (and corresponding group harm) to the conversation surrounding LLMs, which could, in turn, benefit other marginalized groups and their digital representations.¹²¹

Beyond regulatory efforts, Native Nations can use the emergence of LLMs and the corresponding data ecosystems as a chance to further establish their sovereignty along a number of axes. Given that economic development and financial independence are cornerstones of Native Nations' ability to act as sovereigns, data licensing and royalty agreements are one way that Native Nations could directly derive financial benefit from the development of new models.

¹¹⁸ While this paper represents a survey of possibilities, it emphatically does not attempt to represent the variety and depth of tribal perspectives that would be essential to any robust action in this space.

¹¹⁹ "Native American forests and tribal forest management practices have sustained Indigenous communities, economies, and resources for millennia." Michael J. Dockry & Serra J. Hoagland, *A Special Issue of the Journal of Forestry—Tribal Forest Management: Innovations for Sustainable Forest Management*, 115 J. FORESTRY 339, 339 (2017).

¹²⁰ In Washington State, the Swinomish Indian Tribal Community has created a model of health care for treating Substance Use Disorder that centers on traditional care practices. Didg^wálič Aboriginal Language and Culture Centre, DIDG^wÁLIČ, <https://www.didgwalic.com/index.html> (last visited Mar. 19, 2024).

¹²¹ Better practices around child welfare data, for example, stand to reduce structural racism in that particular system.

The Māori¹²² have shown particular initiative in this space, establishing use guidelines for data with built-in requirements that applications be beneficial to tribal interests.¹²³

The development lifecycle for LLMs is another potential site for tribal engagement. Representation is critical to achieving just results in this space, where educational programs in data science,¹²⁴ machine learning, and adjacent disciplines could be pivotal. While the development of new LLMs themselves may remain out of reach in the foreseeable future (with new models costing upwards of \$100 million¹²⁵ to train), having a seat at the table through employment (both through influence and drawing considerable compensation) is an important middle ground.

B. Techniques from Information Theory and Computer Science

There are a number of promising trends in data and information science that could provide elements of a possible solution. Data-labeling practices could allow for selective handling (and exclusion, when deemed necessary) of sensitive data and recognition of different classes of data could be built into the training process itself. An increased recognition of Traditional Knowledge labels¹²⁶ and other structural means of labeling data as having Indigenous import could be the foundation for specialized data handling for this information, potentially keeping it separate in auxiliary models until the knowledge "mixing" that happens in foundation models is better understood. In the long term, as researchers better understand how to measure information provenance in large models, these structural labels could offer a way to create rules-based predicates for restricted sharing, presentation, and generation when considering Indigenous content.

Machine learning as a field has (finally) become increasingly concerned with the presence of bias in models, and now offers a number of unbiasing techniques as well.¹²⁷ However, the most popular methods rely on output checking, which is a late-stage intervention that doesn't address the fundamental issues within the model itself. Under ideal circumstances, bias audits would result in discarding biased models; in practice, they often result in brittle post-

¹²² See Te Mana Raraunga, *supra* note 33.

¹²³ Noting the historically extractive nature of corporations, Māori licenses include a number of built-in protections around the use of tribal cultural property.

¹²⁴ Projects like Seattle's Urban Indian Health Institute are laying the groundwork for this, working to "decolonize data, for Indigenous people, by Indigenous people." *Urban Indian Health Institute*, SEATTLE INDIAN HEALTH BOARD, <https://www.sihb.org/services-and-programs/urban-indian-health-institute/> (last visited Mar. 19, 2024).

¹²⁵ "GPT-4, the latest of those projects, was likely trained using trillions of words of text and many thousands of powerful computer chips. The process cost over \$100 million." Will Knight, *OpenAI's CEO says Age of Giant AI Models Is Already Over*, WIRED (Apr. 17, 2023), <https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/> (last visited Mar. 19, 2024).

¹²⁶ "The TK and BC Labels are an initiative for Indigenous communities and local organizations. ... The TK Labels support the inclusion of local protocols for access and use to cultural heritage that is digitally circulating outside community context's." TK LABELS, <https://localcontexts.org/labels/traditional-knowledge-labels/> (last visited Mar. 19, 2024).

¹²⁷ Ninaren Mehrabi et al., *A survey on bias and fairness in machine learning*, ACM COMPUTING SURVEYS (2019).

filtering that doesn't address the underlying problematic correlations in the model, and instead seeks to prevent it from "saying the quiet thing out loud."

To address bias in models at a deeper level, training data must be filtered by identifying and removing instances of bias. Currently, disclosure of training data is not required by any United States law, although, as noted above, the EU is contemplating the addition of a training data disclosure provision in its upcoming AI Act.¹²⁸ Such a provision would be a tremendous boon to tribal interests, allowing them to understand exactly what has been included, and what data sovereignty provisions might apply. It is also possible to create "synthetic" datasets, where a secondary dataset mirrors the statistical properties of the source dataset while removing sensitive information (like personal identifiers). Synthetic data also allows for intentional re-weighting of training data,¹²⁹ which would complement this approach, creating better models at training time and reducing the fragility of mitigations.¹³⁰

The security and privacy fields also offer a number of relevant techniques. Individual traits can be protected from sharing through required aggregation thresholds, called *k*-anonymity,¹³¹ where results containing fewer than *k* instances are excluded from consideration. One can also consider the selective addition of noise to the output (differential privacy),¹³² which prevents the re-identification of individuals when proper safeguards are observed. By establishing robust standards for the use of these techniques, Native Nations and tribal members would have better tools to participate in broad demographic data projects without the risk of re-identification or the misuse of data at an individual level.¹³³

C. The Role of Non-Tribal Governments

To prevent the adoption of piecemeal solutions, a strong federal presence (with explicit decision-making power reserved for Native Nations) would make a major difference in fortifying IDSov in the face of the growing prevalence of LLMs and related techniques. The federal government has stepped into this role many times in the history of federal Indian law, both for

¹²⁸ European Parliament, *supra* note 117.

¹²⁹ These statistical properties can be manipulated in generating synthetic data, potentially overcoming biases in data gathering techniques (e.g. increasing the representation for an under surveyed population). See Alex Watson, *Reducing AI bias with Synthetic Data*, GRETEL.AI (Jan. 11, 2021), <https://gretel.ai/blog/reducing-ai-bias-with-synthetic-data> (last visited Mar. 19, 2024).

¹³⁰ Unsurprisingly, the earlier in the process at which bias is addressed, the more comprehensively it can be mitigated.

¹³¹ See Latanya Sweeney, *k-anonymity: a model for protecting privacy*, 10 INT'L J. ON UNCERTAINTY, FUZZINESS AND KNOWLEDGE-BASED SYS. (2002).

¹³² Cynthia Dwork, *Differential Privacy*, INT'L COLLOQUIUM ON AUTOMATA, LANGUAGES AND PROGRAMMING (2006).

¹³³ To be clear, there is an inherent tension between privacy and specificity here, one that is pronounced for smaller populations. Techniques like differential privacy add synthetic noise to data, reducing its accuracy (and the smaller the population, the more noise must be added).

better (e.g. ICWA,¹³⁴ some IRA provisions) and worse (allotment, termination).¹³⁵ Although state governments could attempt to step in,¹³⁶ the federal government's responsibility to Native Nations provides a much clearer basis for federal action,¹³⁷ as well as broader impact.

There is precedent for federal recognition of cultural artifacts through NAGPRA,¹³⁸ and a similar approach could be relevant here, with civil and criminal penalties for misrepresentation or theft of tribal cultural property (whether tangible or intangible). As a backstop against the possibility of digital erasure, Congress could compel LLM creators to engage in good faith bargaining efforts with Native Nations (similar to IGRA).¹³⁹ Unless the creators are States, this compulsion would not be subject to the kind of sovereign immunity challenge as seen in *Seminole*.¹⁴⁰

Absent federal action (or as a complement to it), Native Nations could also work in coalition to develop a common set of legal standards for LLM adoption of tribal data and information.¹⁴¹ While the loss of preferences for individual Native Nations may render this approach ultimately infeasible, this could similarly forestall the digital erasure stemming from model creators deciding that tribal jurisdiction is simply too much trouble and to avoid the problem altogether. TK labels and related work provide a possible starting point for such efforts.

VI. CONCLUSION

In sum, LLMs present a complex mix of dangers (and select opportunities) for Indigenous Data Sovereignty and Indigenous economic development more broadly. The current path for these models will reproduce and amplify the harms that are already present in many modern information systems. Given the enormous power wielded by corporations and their advocates in the current race to create more powerful LLMs, a coordinated countervailing response from Native Nations and the federal government is urgently needed. Although any such efforts should

¹³⁴ ICWA in particular is notable for its delegation of decision-making power to Native Nations themselves, giving tribal courts jurisdiction over adoption proceedings. Oftentimes, the most effective thing that the federal government can do is create enough legal space (and return stolen resources) for Native Nations to exercise the sovereignty that they have practiced since time immemorial.

¹³⁵ This see-sawing of fortunes understandably generates mistrust around federally-backed solutions. Any federal solution should endeavor to transfer power to Native Nations as durably as possible, rather than render it subject to the caprices of future federal lawmakers.

¹³⁶ Potentially a fraught course of action, given that states have been called the "deadliest enemies" of Indian people, though Professor Fletcher suggests refining this frame. Matthew Fletcher, *Retiring the "Deadliest Enemies" Model of Tribal-State Relations*, 43 TULSA L. REV. (2007).

¹³⁷ While not legally foolproof, the Court has generally upheld that the federal government has a trust relationship to Native Nations and may enact provisions that are demonstrably in the interest of those Native Nations. To be clear, the federal government has almost always made this determination (of what is "in the interest" of Native Nations) on its own, thus the need for the centering of tribal leaders this time around. *See United States v. Mitchell*, 445 U.S. 535, 100 S. Ct. 1349, 63 L. Ed. 2d 607 (1980).

¹³⁸ Native American Graves Protection and Repatriation Act (NAGPRA), 25 U.S.C. § 3001-3013 (1990).

¹³⁹ Indian Gaming Regulatory Act (IGRA), 25 U.S.C. § 2701-2721 (1988).

¹⁴⁰ *Seminole Tribe v. Florida*, 517 U.S. 44 (1996).

¹⁴¹ *See*, for example, the Native American Rights Fund and work like THE IMPLEMENTATION PROJECT, <https://narf.org/> <https://un-declaration.narf.org/> (last accessed Mar. 27, 2024).

be led by Native Nations themselves, there are important places where the federal (and potentially state) government(s) can support these efforts. With thoughtful and careful efforts, LLMs can avoid perpetuating centuries-old harms, and instead can prioritize centering Indigenous knowledge and information. This approach not only supports Indigenous communities, it promotes more inclusive and equitable systems for everyone.