My PhD has focused on how to align sociotechnical systems with human values in ways that are effective and accountable to the people they impact. I work at the intersection of human-computer interaction (HCI) and responsible AI to understand the challenges of designing and deploying systems in socially complex settings. Although technology increasingly shapes critical aspects of public life, there remain many challenges when systems are deployed in new contexts. From matching algorithms to generative AI, systems that appear successful in evaluations can, in practice, produce behaviors that frustrate, marginalize, or misrepresent stakeholders [5, 18]. As sociotechnical systems increasingly determine who accesses critical resources and influence how we understand our world, the gap between our intended values versus what gets enacted grows more concerning.

Values alignment remains difficult for various reasons. Values, like fairness or justice, cannot be directly measured with instruments. Furthermore, they can have multiple, conflicting interpretations—shaped by race, class, and lived experience. When designing systems, determining which values to enact entails normative decisions about whose interpretations to prioritize. As systems increasingly become part of public life (e.g., policy, decision-making, information seeking), they must also establish trust and legitimacy across affected stakeholders—who may have diverse values and needs. My work has shown that systems that encode assumptions from one context break down when deployed into new settings [10]. **Alignment requires attention to local values, stakeholder needs, and context—yet redesigning a system for every setting is neither practical nor sufficient, as we cannot fully anticipate how contexts will shape its use. This leads me to ask: how can stakeholders participate in shaping how systems behave?**

To answer this question, my research approach combines empirical mixed-methods studies, participatory design, and systems-building. I partner with affected stakeholders to understand how systems are used, what they need, and where tensions emerge. I map technical design choices to the social phenomena they produce—analyzing how system structure and deployment contexts mutually shape one another. From these analyses, I develop interventions and evaluation methods that better align systems with stakeholder values. **My approach treats alignment not as something done to systems once, but as a process stakeholders actively participate in through transparency, control, and recourse**.

I. Evaluative Frameworks for Algorithmic Legitimacy

In 2011, the San Francisco Unified School District launched an algorithmic assignment system to promote equity and racial integration for 57,000 students, but by 2018, schools were more segregated. The district grappled with how to operationalize diversity in zone boundaries, while still offering families proximity and predictability. The underlying algorithm was optimized for preference satisfaction, not diversity, and legal constraints precluded measuring racial outcomes directly. Without a way to evaluate the system or assess whether parents would accept their assignments, the district risked losing legitimacy—and families. Parents were already fleeing to private schools, contributing to a wave of school closures across the district. How can we evaluate policy-facing algorithms under conditions of values contestation and legal ambiguity?

Procedural justice theory holds that when people judge the legitimacy of decisions, they often care more about how they experienced the process than the outcome itself [18]. Crucially, the theory provides validated measures that predict whether stakeholders will accept outcomes and view decision-makers as legitimate. This offered a path forward: procedural justice could let SFUSD evaluate its system's legitimacy without measuring outcomes directly—essential when legal constraints preclude race-based metrics. But would procedural justice actually predict outcome acceptance in this context? And what would a fair process look like to parents?

In 2022, I collaborated with a legal scholar to find out. We surveyed 1,171 SFUSD parents, measuring procedural justice antecedents, perceptions of fairness, and collecting open-ended fairness definitions. The results showed that

procedural justice strongly predicts outcome acceptance: parents who viewed the assignment process as fair were more likely to judge their outcome as fair, even when it was not ideal. This offered a path forward—a framework for evaluating the assignment process under legal constraints that could actually measure whether families were likely to accept their assignments. I also found that parents' definitions of fairness revealed systematic differences by race and socioeconomic status, and how the algorithm was communicated shaped parents' overall expectations of fairness (e.g., parents defined fairness as “receiving their top choice”) [8].

These findings revealed a deeper tension: fairness judgments are socially situated, shaped by neoliberal logics that recast fairness as choice satisfaction [1, 2, 4]. Parents framed diversity as benefiting their child rather than as collective equity, and many defined fairness as "going to their neighborhood school"—echoing the legacy of racial segregation SFUSD aims to dismantle. This raised three questions: (1) How does cultural capital shape parents' responses to algorithmic assignment? (2) How do prior expectations influence perceptions of the process? (3) Can reframing communication about the algorithm shift perceptions of fairness?

II. Measuring Fairness and Legitimacy in Algorithmic Social Policy

I conducted 25 follow-up interviews with SFUSD parents, investigating how cultural capital, prior educational experiences, and political values shaped perceptions. Higher-SES parents, including many who received the best outcomes, were often most critical of SFUSD. Many framed equity provisions as "holding the top down," arguing that retaining affluent families should itself be a goal since their participation sustained school resources. Other parents accepted redistribution as legitimate and described their outcomes in collective rather than individual terms. Resistance to the algorithm reflected which political values parents believed the system should embody.

To test whether communication could shift fairness perceptions, I ran an online RCT with U.S. parents (N = 409) in a simulated school assignment process. After receiving an unfavorable assignment, parents got one of three explanations: none, an XAI-oriented explanation (how the algorithm produced the decision), or a collective equity explanation (why the outcome served broader equity goals). Both explanations meaningfully improved perceptions of fairness ($\beta \approx 0.52^{***}$), trust ($\beta \approx 0.46^{*}$), and legitimacy ($\beta \approx 0.60^{**}$). The collective framing matched the XAI condition on these measures and went further: parents reported stronger support for expanding the algorithm ($\beta = 0.36^{*}$) and weaker support for eliminating it ($\beta = -0.43^{**}$). Taken together, these studies show that evaluating a sociotechnical system ecologically—across stakeholder positions and against the political logics it encodes—reveals that preferences are not stable, and that what counts as a good outcome depends on which values the system is framed as pursuing. This work was submitted to CHI 2027 [9].

Taken together, these studies suggest that fairness and acceptance in public-sector sociotechnical systems require a critical account of race, class, position, and power: not only who stakeholders are and what they say they prefer, but how the system’s rules and justifications interact with—and can reproduce—existing relations of dominance [9].

III. Developing Measurement Practices for Evaluating Generative AI Systems

Evaluations of generative AI systems shape high-stakes decisions: which models get deployed, what harms get flagged, whose concerns become legible to developers and policymakers. Yet the concepts these evaluations target—capabilities like "reasoning," harms like "stereotyping"—remain contested across use cases, cultures, and communities. Too often, researchers move directly from vague concepts to evaluation tools, embedding definitional choices in technical artifacts that foreclose participation from those who lack technical fluency but understand real-world stakes. Measurement theory for GenAI evaluation addresses this by first developing systematized concepts—precise specifications of a concept's definition, components, and boundaries [19]—before building evaluation tools. Because systematized concepts impact every downstream stage of measurement, engaging stakeholders at this stage gives them a chance to advocate for understandings that reflect their priorities [3, 19]. Yet

few examples exist of how this can be done [6]. **I asked, how can participatory methods integrate stakeholder perspectives into AI evaluation, and what does this approach make visible?**

Working with Microsoft Research, I explored how to involve stakeholders in validating and refining systematized concepts used to evaluate generative AI systems [22]. I conducted six in-person participatory design workshops with 23 participants to validate and refine an existing systematized concept of "erasure"—speech that disempowers groups and individuals by erasing their differences, resulting in entrenching harmful social hierarchies [5]. Participants surfaced gaps the concept missed (e.g., failures to credit cultural contributions, delegitimizing labels like "woke"), revealed tensions around which groups and hierarchies should be included, and distinguished between instance-level erasure (a single harmful output) and distributional erasure (harm that accumulates through repeated omission across many outputs)—for example, a GenAI system that only produces heterosexual romance stories. This approach surfaced gaps between participants' lived experiences and what the systematized concept captures, revealing the breadth of alternative definitions that could have been developed. Beyond validation, this work demonstrates how stakeholders can participate at distinct stages—concept development, validation, and tool design—informing evaluation tools that scale without requiring end-to-end engagement at every deployment [11].

Yet measurement frameworks are only one lever for shaping the policies and guardrails that govern system behavior. Equally important is building interaction-level infrastructure that exposes how outputs are produced and enables users to steer model behavior. **Together, these approaches treat alignment not just as something done to models, but as something users actively participate in through transparency, control, and recourse.**

Future Research Agenda

Sociotechnical systems are embedded in rich, complicated social contexts. This situatedness makes it difficult for "scalable," "general-purpose" systems to behave desirably in new deployment contexts—a mismatch that can create real harm in the world [7, 10, 17]. My research demonstrates this mismatch and provides models for translating "thick descriptions" [16, 21]—local histories, resource constraints, competing values, political specificities—into evaluative frameworks, participatory methods, and policy recommendations. While there is no one-size-fits-all approach, a deep understanding of deployment contexts is necessary for systems that are accountable, effective, and reflective of values. In particular, race, class, intersectionality, and local politics are constitutive of how systems are experienced and trusted—foregrounding how identity is entangled with algorithmic systems.

From thick descriptions to thicker evaluations [12], I aim to design interventions and measurement practices flexible and accessible enough to preserve social complexity rather than flatten it. **In particular, I do this by examining how institutions and infrastructures configure sociotechnical systems, and are reconfigured by them; how community engagement can shape AI governance; and what it means for generative AI to serve public responsibilities rather than evade them.**

I) Infrastructures that support community engagement in the governance and reconfiguration of AI systems

Evaluation increasingly steers deployment and policy decisions, but many measurements are not yet valid or robust to real-world use. Many risky model behaviors are difficult to surface in pre-deployment testing, motivating new measurement approaches that can detect these failure modes. My vision is to develop a third-party evaluation infrastructure that is (1) participatory—grounded in impacted stakeholders' definitions of risk and capability, (2) iterative—updated as deployment contexts shift, and (3) auditable—with explicit rubrics and monitoring practices that support reproducibility and trust. Going forward, I want to make these interventions more accessible by building mechanisms for continuous feedback: open evaluation artifacts (datasets, rubrics, test protocols) that researchers and public experts can critique, revise, and extend.

II) Accountability in Algorithmic Decision-Making and GenAI

In future work, I will extend my interdisciplinary approach to settings where algorithmic systems influence people's opportunities and access to resources (e.g, allocative harms, quality of service, and social system harms). GenAI poses a poor regulatory target because of its generality and lack of task specification. This is further complicated by a distributed value chain, where liability is difficult to pinpoint (e.g., multiple actors may develop, finetune, or integrate models into new products)—when an AI system causes harm, it is difficult to pinpoint which part of the chain is legally responsible. **My guiding question is: how can we build accountability into algorithmic decision-making systems that increasingly govern public life?**

I will work with policymakers and stakeholders to clarify regulatory targets around applications and consequences. This means directing regulatory attention toward deployment context and intended use, so scrutiny can be calibrated to application-specific risks. Methodologically, I will trace how harms are conceived, produced, and realized across an ecosystem, identifying the structural dependencies that enable harm and the points where responsibility is missing or diffused. The goal is to produce actionable governance tools—clearer lines of responsibility, context-sensitive evaluation requirements, and mechanisms for contestability and recourse—that make accountability feasible under real-world conditions of distributed deployment.

Furthermore, I imagine exploring this by incorporating high-level societal objectives into operational specifications in algorithm design. For example, what does it mean for a system to have a notion of "justice" when allocating resources, while remaining explicit about trade-offs, contestation, and legal constraints? I also hope to examine whether learning-from-feedback (e.g., RLHF-style preference aggregation) approaches can support governance decisions, and under what conditions they amplify dominant preferences or obscure accountability. The goal is to develop empirically grounded methods from other disciplines, such as sociology and law, for translating contested values into decision criteria.

III) Reliable, Explainable, and User-Controllable AI Interaction

I am interested in bringing lessons from my work on explanations in algorithmic decision-making to responsible AI and large-language models (LLMs). This work is important due to two present dynamics. First, millions of people rely on LLMs for high-stakes information (e.g., health, science, and current events), and organizations increasingly expect employees to use these systems to work more efficiently. Second, LLM performance is unreliable and interaction-dependent: outputs can be subtly inaccurate, and the quality users receive varies with their ability to elicit, verify, and revise answers across turns. Together, these dynamics create three risks: (i) inaccurate responses can misinform consequential decisions, (ii) fluent, authoritative language can produce miscalibrated trustand (iii) institutions may overestimate what these systems can do and shift the burden of failure onto end users (e.g., treating "prompting skill" as a job requirement), leading to unfair expectations and unequal access. **I ask, how can we build more reliable and explainable human-AI interactions? What design interactions can we leverage for increasing user control, transparency, and critical inquiry?**

Already, with collaborators such as Gary Hsieh and Lucy Lu Wang at the University of Washington, I aim to build and empirically evaluate user-facing tools that support meaningful control over AI behavior and treat users as active collaborators, rather than passive recipients. One idea is to allow users to select spans of text and back-attribute parts of the model's output to an editable decision stack (e.g., reasoning traces, alternative output choices) produced during turn-taking. While current consumer-facing interfaces increasingly expose summaries of model reasoning (e.g., ChatGPT; likely to mitigate model distillation), or offer span-level editing actions like "Improve" or "Explain" (e.g., Opus), these interactions do not let users revise upstream decisions in a way that reliably propagates through the whole response. In my proposed system, users can edit specific spans in the decision stack, and the system will regenerate downstream content while preserving unaffected parts.

Citations

[1] John Ambrosio. 2013. Changing the Subject: Neoliberalism and Accountability in Public Education. *Educ. Stud.* 49, 4 (July 2013), 316–333. https://doi.org/10.1080/00131946.2013.783835

[2] Lawrence Angus. 2015. School choice: neoliberal education policy and imagined futures. *Br. J. Sociol. Educ.* 36, 3 (April 2015), 395–413. https://doi.org/10.1080/01425692.2013.823835

[3] Alexandra Chouldechova, Chad Atalla, Solon Barocas, A. Feder Cooper, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Matthew Vogel, Hannah Washington, and Hanna Wallach. 2024. A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. https://doi.org/10.48550/arXiv.2412.01934

[4] Christina Convertino. 2017. State disinvestment, technologies of choice and 'fitting in': neoliberal transformations in US public education. *J. Educ. Policy* 32, 6 (November 2017), 832–854. https://doi.org/10.1080/02680939.2017.1324113

[5] Emily Corvi, Hannah Washington, Stefanie Reed, Chad Atalla, Alexandra Chouldechova, P. Alex Dow, Jean Garcia-Gathright, Nicholas J Pangakis, Emily Sheng, Dan Vann, Matthew Vogel, and Hanna Wallach. 2025. Taxonomizing Representational Harms using Speech Act Theory. In *Findings of the Association for Computational Linguistics: ACL 2025*, July 2025. Association for Computational Linguistics, Vienna, Austria, 3907–3932. https://doi.org/10.18653/v1/2025.findings-acl.202

[6] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (*EAAMO '23*), October 30, 2023. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3617694.3623261

[7] Alex Hanna and Tina M. Park. 2020. Against Scale: Provocations and Resistances to Scale Thinking. https://doi.org/10.48550/arXiv.2010.08850

[8] Tonya Nguyen, Sabriya Alam, Cathy Hu, Catherine Albiston, and Niloufar Salehi. 2024. Definitions of Fairness Differ Across Socioeconomic Groups & Shape Perceptions of Algorithmic Decisions. *Proc ACM Hum-Comput Interact* 8, CSCW2 (November 2024), 519:1-519:31. https://doi.org/10.1145/3687058

[9] Tonya Nguyen, Cathy Hu, Liza Gak, Serene Cheon, Tara Kaviani, Catherine Albiston, and Niloufar Salehi. 2026. How Explanations and Political Frames Shape Perceptions of Algorithmic Social Policies. Working Paper, Submitted to CHI 2027.

[10] Tonya Nguyen, Darya Kaviani, and Niloufar Salehi. "It Actually Doesn't Feel Very Mutual:" How Technology Impacts the Values of Mutual Aid Groups in Practice. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, Yokohama, Japan.

[11] Nguyen, Tonya, Jean Garcia-Gathright, Alexandra Chouldechova, Hannah Washington, Hanna Wallach, Jennifer Wortman Vaughan. 2026. "Validating and Refining Generative AI Evaluations via Stakeholder Engagement." In Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26). Montreal, Canada.

[12] Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025. The Case for "Thick Evaluations" of Cultural Representation in AI. https://doi.org/10.48550/arXiv.2503.19075

[13] Samantha Robertson, Tonya Nguyen, Cathy Hu, Catherine Albiston, Afshin Nikzad, and Niloufar Salehi. 2023. Expressiveness, Cost, and Collectivism: How the Design of Preference Languages Shapes Participation in Algorithmic Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (*CHI '23*), April 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3580996

[14] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2021. Modeling Assumptions Clash with the Real World: Transparency, Equity, and Community Challenges for Student Assignment Algorithms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (*CHI '21*), May 07, 2021. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445748

[15] Samantha Robertson, Tonya Nguyen, and Niloufar Salehi. 2022. Not Another School Resource Map: Meeting Underserved Families' Information Needs Requires Trusting Relationships and Personalized Care. *Proc ACM Hum-Comput Interact* 6, CSCW2 (November 2022), 316:1-316:23. https://doi.org/10.1145/3555207

[16] Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. Culture: Critical Concepts in Sociology, 173-196.

[17] Anna Lowenhaupt Tsing. 2012. On Nonscalability. *Common Knowl.* 18, 3 (August 2012), 505–524. https://doi.org/10.1215/0961754X-1630424

[18] Tom R. Tyler. 1988. What Is Procedural Justice - Criteria Used by Citizens to Assess the Fairness of Legal Procedures. *Law Soc. Rev.* 22, (1988), 103.

[19] Hanna Wallach, Meera Desai, Nicholas Pangakis, A. Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2024. Evaluating Generative AI Systems is a Social Science Measurement Challenge. https://doi.org/10.48550/arXiv.2411.10939

[20] 2020. Board Policy 5101.2 Elementary School Student Assignment. Retrieved April 11, 2021 from https://go.boarddocs.com/ca/sfusd/Board.nsf/files/BVYUGB7BF68F/$file/BP%205101.2%2C%20Elementary%20School%20Student%20Assignment.pdf

[21] Alexandrova, A., Fabian, M. Democratising Measurement: or Why Thick Concepts Call for Coproduction. Euro Jnl Phil Sci 12, 7 (2022). https://doi.org/10.1007/s13194-021-00437-7