RESEARCH STATEMENT

MICHAEL BERNSTEIN

Computing platforms mediate our interactions with each other: how we work, how we socialize, and even how we govern. By changing these computing platforms, we change human social behavior: as per Winston Churchill, "We shape our buildings; thereafter they shape us." How do we design these social computing systems to mitigate their harms and amplify their benefits? My research in human-computer interaction advances systems, designs, and models for social computing systems. In doing so, I aim to shed light on long-standing research questions: How do we cross the gap between social scientific knowledge and concrete system designs that manifest this knowledge, to facilitate pro-social interactions and collaborations? How can we model human attitudes and behavior to help us craft these designs and social policy? How should we design our computational tools if we want them to be successfully embedded into society?

ENABLING INTERACTIVE SIMULACRA OF HUMAN BEHAVIOR. Every social system makes assumptions about how design shapes behavior. For a social network, implementing a feature might be assumed to reduce the spread of misinformation; for a policymaker, implementing a new law might be assumed to increase small business investment. Are these assumptions accurate? Behavioral simulation has arisen as a key tool powering both science and design by enabling "what if?" counterfactuals, helping policymakers recognize unexpected consequences of an intervention, and empowering tighter, more iterative reflective loops. However, existing simulation models have remained too brittle to accommodate the complex contingencies of human behavior.

Recently, we developed generative agents [Park et al. UIST 2023], computational software agents that draw on generative AI models to enable believable simulation of human behavior. Underlying generative agents is an architecture that enables memory of the agent's interactions, reflection to generalize from experiences to higher-level observations, and long-term planning. To demonstrate these agents, we populated an interactive sandbox environment inspired by The Sims with a small town of twenty-five generative agents. In an evaluation, these generative agents produce believable individual and emergent social behaviors: for example, starting with only a single user-specified notion that one agent wants to throw a Valentine's Day party, the agents autonomously spread invitations to the party over the next two days, make new acquaintances, ask each other out on dates to the party, and coordinate to show up for the party together.

Generative agents point toward the possibilities of societal simulations, but they also offer utility in individual end-user tools. For example, we have developed agents that simulate the kinds of anti-social behaviors that might later arise in an online community [Park et al. UIST 2022], enabling social media designers to plan before problems arise. We also have created agents that support skill training by becoming simulated interlocutors for rehearsing difficult interpersonal conflicts [Shaikh et al. CHI 2024]. Extending from models of behavior to models of attitudes, we developed an agent that learned how to integrate into an online social network by observing which of its questions people are willing to answer about their photo posts, leading to interactions with over 236,000 people on a photo sharing social network [Krishna et al. PNAS 2022].

This work has brokered substantial interest in multi-agent simulation and behavioral simulation. The generative agents research has received over 500 citations within a year of release, and won a Best Paper Award at UIST 2024; it was covered by media including Nature, NBC, and it led to an invited talk at TED AI; its open-source GitHub repository has over 1,700 forks and nearly 15,000 stars. The conflict rehearsal environment project is now in active collaboration with a civil society organization that has the capacity to potentially deploy it to over a half million students. Moving forward, we seek to develop a scientific method of generative agent-based modeling: how accurate are these simulations, under what conditions, and what scientific method enables rigorous theory testing?

SOCIAL ENVIRONMENTS SHAPE HUMAN-AI INTERACTION. Developing AIs without attending to social context yields embarrassing errors. While AI traditionally focused on tasks with clearly correct or incorrect answers, in social computing we instead face tasks ranging from online toxicity to misinformation detection where it is far more likely that different groups have resilient, strong disagreement about the correct answer [Gordon et al. CHI 2021]. Keeping only the highest-voted annotator response unfortunately erases minority viewpoints. In response, we introduced Jury Learning [Gordon et al. CHI 2022], which instead draws on the metaphor of jury selection to take the perspective that the developer ought to explicitly directly specify which groups' opinions, in what proportion, should determine the AI's behavior. For example, for an online toxicity model used in social media moderation, the developer might want those commonly impacted by online harassment to feature centrally. The jury learning algorithm integrates deep learning text classification models with recommender system architectures to estimate each juror's opinion, then aggregates those

jurors' points of view into a prediction. A field evaluation with Jury Learning finds that practitioners construct diverse juries whose resulting algorithms alter 14% of classification outcomes compared to models that ignore disagreement.

Anchoring the development of AI in social contexts yields research questions that are crisp and answerable while maintaining strong generalizability. For example, why are AI errors in social contexts such as content moderation and worker evaluation seen as so problematic? Much like "street-level bureaucrats" such as police and judges, AIs translate high-level social policy into low-level decisions. We developed by analogy a theory of street-level algorithms [Alkhatib and Bernstein CHI 2019], arguing that AIs create negative outcomes when faced with novel situations where they cannot, unlike street-level bureaucrats, dynamically reflect on and refine their decision criteria. We also observed that the conceptual metaphors projected by AI systems (e.g., "personal butler") causally impact peoples' usage intentions—with metaphors that signal high competence often backfiring [Khadpe et al. CHI 2020]. Unsurprisingly, then, on social media moderation tasks, people view expert panels as carrying more procedural legitimacy than algorithms [Pan et al. CSCW 2022]. However, we observed encouraging evidence that people do make strategic decisions about when to rely on AI explanations in decision-making [Vasconcelos et al. CSCW 2023]. In response, we moved to allow end users and communities to directly participate in authoring [Lam et al. CHI 2023] and auditing models [Lam et al. CSCW 2022], as well as engaging in computational social science investigations [Lam et al. CHI 2024; Cao et al. CHI 2023].

This work on socially-informed human-AI interaction received Best Paper awards at CHI 2023, CHI 2022, and CHI 2019, and was honored with the Computer History Museum Patrick J. McGovern Tech for Humanity Prize. Our future work will complement these efforts by developing end-user and community tools for rapid AI customization.

COMPUTATION SHAPES CROWDSOURCING AND SOCIAL MEDIA. In our earlier work, we demonstrated that computational tools can power large-scale crowdsourcing collaborations. We began by establishing crowd programming patterns such as Find-Fix-Verify and interactive applications of crowds as in word processing [Bernstein et al. UIST 2010], then demonstrated that computation can coordinate experts into on-demand flash teams [Retelny et al. UIST 2014] and flash organizations [Valentine et al. CHI 2017] that tackle complex goals such as product design, software engineering, and game production. Our work served as a rallying cry for a pro-social future of work [Kittur et al. CSCW 2013], demonstrating the first platforms for online gig worker collective action [Salehi et al, CHI 2015]. The research also earned a UIST Lasting Impact Award and Best Paper Awards at UIST 2010, UIST 2014, and CHI 2017; was used to scale the ImageNet Challenge [Russakovsky et al. IJCV 2015] and Visual Genome [Krishna et al. IJCV 2017] datasets; and forms the basis of a Flash Teams book [Valentine and Bernstein, forthcoming from MIT Press].

Today, online social systems remain troubled, and I believe that we are falling prey to failures of imagination in how to improve them. One in twenty posts that remain on popular social media platforms violate the platforms' own rules [Park et al. CSCW 2022], and much antisocial content arises from regular users, not inveterate trolls [Cheng et al. CSCW 2017]. In response, we recently developed the means for social media ranking algorithms to augment their traditional individualist values (e.g., engagement) with societal values [Jia and Lam et al. CSCW 2024; Bernstein et al. JOTS 2023]. To do so, we adapt construct measurements from the behavioral sciences, whose conceptual and linguistic precision is a strong fit for large language models. Our model built around the construct of anti-democratic attitudes, for example, achieves high levels of agreement with expert annotators; across a series of experiments, integrating this model into social media ranking achieves bipartisan reductions in partisan animosity. Looking from the algorithms to the platform design, we introduced the Form-From design space of social media systems to aid designers in identifying the consequences of their design choices [Zhang et al. CSCW 2024]. Expanding this design palette required developing tools for custom governance structures on communities such as Reddit and Discord, enabling everything from regular elections to peer juries [Zhang, Hugh and Bernstein UIST 2020]. Finally, to help explore new alternative models for social media, we developed algorithmic curation algorithms that can help a community or platform stay true to a north star of specific curators' tastes, while still leveraging the input from the larger group [He et al. CSCW 2023].

This work was awarded Best Paper at CSCW 2023. In future work, we will continue developing novel social media designs, with the goal of offering alternative modes that retain the prosocial benefits of social interaction while addressing many of current designs' negative impacts by design, rather than by band-aid.

CONCLUSION. Whether at the small scale of pairs or teams, or at the large scale of crowds and societies, my core interest is in crafting new socio-technical arrangements: developing new design, empirical, and technical tools that help us facilitate better social interactions with each other.