GOOD ENOUGH EXPLANATIONS: HOW CAN LOCAL PUBLICS UNDERSTAND AND EXPLAIN CIVIC PREDICTIVE SYSTEMS?

A Dissertation
Presented to
The Academic Faculty

By

Shubhangi Gupta

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Literature, Media, and Communication
Department of Digital Media

Georgia Institute of Technology

August 2024

GOOD ENOUGH EXPLANATIONS: HOW CAN LOCAL PUBLICS UNDERSTAND AND EXPLAIN CIVIC PREDICTIVE SYSTEMS?

Thesis committee:

Dr. Yanni Loukissas, Advisor
School of Literature, Media, and Communication
Georgia Institute of Technology

Dr. Chris Le Dantec
College of Arts, Media, and Design and Khoury College of Computer Sciences
Northeastern University

Dr. Noura Howell
School of Literature, Media, and Communication
Georgia Institute of Technology

Dr. Lauren Klein
Departments of English and Quantitative Theory & Methods
Emory University

Dr. Richmond Wong
School of Literature, Media, and Communication
Georgia Institute of Technology

Date approved: Jul 10, 2024

ACKNOWLEDGMENTS

I feel privileged to have been surrounded by very kind, supportive, and joyful people during the course of my PhD. They have helped me grow not just as a researcher, but as a person. Here, I want to express my heart-felt gratitude to everyone without whom this dissertation would not be possible.

To my advisor, Yanni, your confidence in my work has meant more to me than I can put in words here. Thank you for being the mentor, researcher, collaborator, and writer I aspire to be one day. You have constantly guided me, challenged me, and inspired me in critical and constructive ways. It has been an absolute honor being advised by you.

To my committee members, Richmond, Noura, Lauren, and Chris, thank you for your support these past few years. You have been very generous with your time and have helped me develop my ideas and writings that have eventually shaped this work in fundamental ways.

To everyone who funded me in this journey; Carol, thank you for inviting me to join the Women, Science, and Technology Learning Community, and financially supporting the entirety of my PhD, never wavering. You have been a constant calming presence in this journey that was sometimes very messy and overwhelming. And Allen, who funded my summers and supported my conference travels, thank you so much.

To my internship mentors, Sonam, Juanis, and Sara, thank you for keeping me in touch with the world outside of academia. You all have been invaluable in helping me discover what I truly want out of my PhD. You all continue to be my mentors to this day and for that, I am so grateful.

To my small but mighty Digital Media family, a lot of us started our grad school during COVID, unaware of what a 'normal' PhD life looks like. We together created a new normal for us, a normal that I have greatly enjoyed and that I will miss so much. Pooja, Mohsin, and Sylvia thank you for constantly brainstorming with me, reviewing my writing, cheering

me on, pushing me to reach for things I thought I could not achieve, and most importantly for being my friend. It would have been a very lonely journey if it weren't for you all.

To all other friends within and beyond Georgia Tech, Kiran, Jaya, Nikhil, Sushant, Aditya, Vishal, Manoj, Atefeh, Vaibhav, and so so many more people who I may not be able to name here, thank you for listening to my many many rants, for helping me disconnect from work when needed, and for making sure I enjoy this journey.

To my fiancé, Nirbhay, you challenged me to start this journey and held my hand as I completed it. Intellectually, you have been my sounding board, my fresh eyes, who has helped me develop each and every argument I make in this work. Emotionally, you have been my rock, who has kept me in touch with who I want to be and what I want to do in this fleeting life. This journey wouldn't be nearly as exciting or gratifying if it weren't for you.

And lastly, my family, my parents Prabha and Puneet, my brother Manush, and my sister (in law) Priyanshi; Mom and Dad, the list of things I should thank you for is endless. You have shown me what it means to live a wholesome life, a happy life, a life full of love and laughter. It is a privilege to be your daughter. I owe it all to you both. Manush, my big brother, whose footsteps I first followed but who helped me become my own person in time. Thank you for being there for me in times that truly mattered. And Priyanshi, my new sister, thank you for joining our family and ending my PhD journey on a high!

Mentors, friends, and family, I will cherish my memories with you all forever. Thank you for everything!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

How can the Explainable Artificial Intelligence (XAI) community support public understanding of the spatial workings and effects of civic predictive systems? Civic processes in the urban smart city are increasingly being governed by automated predictive systems using machine learning models. Despite their widespread use in everyday domains such as education, policing, social services, and economic investments, they continue to remain invisible and inaccessible to local publics, who bear the burden of their effects. XAI and Artificial Intelligence (AI) transparency researchers are increasingly calling for the development of public-centered AI explanations. However, in the context of civic AI, existing techniques fall short in (1) how they understand the consumers and creators of explanations, (2) how they explain the socio-technical assemblages that give rise to AI systems, (3) how they design interactions to create and deliver explanations, and (4) how they conceptualize explanation goals in relation to public action. This dissertation engages in qualitative, participatory, and design-based research to introduce the concept of 'good enough explanations' in response to these challenges. Good enough explanations may not be complete or universal. Instead, as this dissertation formulates, such explanations consist of *ongoing processes* that allow *diverse publics* to *partially* engage with *features of predictive systems* and *assess such systems in relation to their communities*. This dissertation (1) theorizes qualities underlying good enough explanations, (2) engages in the development of such explanations with diverse publics, and (3) suggests theories and strategies to guide the development of systems for good enough explaining.

Ultimately, this dissertation hopes to serve as a guide for XAI researchers, civic organizations, as well as policymakers, as they work together to engage with publics for the democratic oversight, assessment, and regulation of civic AI systems.

CHAPTER 1
INTRODUCTION

How can the Explainable AI (XAI) community support public understanding of the spatial workings and effects of civic predictive systems?

Civic predictive systems, or systems that affect civic processes, are becoming an inseparable part of the urban smart city. They inform several fundamental decisions that govern our everyday lives: who gets parole, bail [1], or economic and social aid [2]; who gets hired [3]; how are students assigned to public schools [4]; how well are teachers performing [5]; which child welfare calls need immediate attention [6, 7]; how un-housed people receive housing resources [8]; who gets a loan application approved [9]; or which routes we, or our public transports, take in a city [10, 11]. They are pervasive.

The use of predictive systems has been presented as making civic processes more efficient, accurate, and objective. They can effectively and quickly process large amounts of data about a variety of subjects and their environments. They can identify patterns not previously known while attempting to overcome biases in human-based decision-making [12]. The hope is that the use of such systems can address the resource constraints within the public sector by off-handing or automating labor, thereby reducing costs and increasing resource allocation capacities [13].

Despite their proposed vision and goals, several civic predictive systems have been shown to cause societal harm and discrimination [14]. There are countless examples of bias [15, 16], disregard of civil rights and liberties [17, 18], or privacy invasions [19]. The centralized and authoritative nature of algorithms gives them the power to create or reinforce unjust (and often invisible) societal structures and inequalities [20].

Geospatial civic algorithms specifically, that make predictions about spaces in the smart city, attempt to fit highly complex cities into neat structures demanded for computation

[21]. Critical geographers, such as Schwanenand and Kwan, demonstrate how social discrimination and marginalization are inherently entangled with space [22]. The overly simplistic mathematical representation of cities creates the kinds of invisible inequalities and incremental injustices that produce large-scale societal effects, such as gentrification and segregation [23]. Numerous cases exemplify the inequitable spatial effects on diverse communities:

> Planning algorithms such as the Market Value Analysis present governing bodies with a "data-driven objective" means of distributing public resources by classifying and segregating communities through the construction of color-coded boundaries. Such organization of space, as Safransky argues, standardizes unjust classification systems and restructures the public sphere in ways that favor some neighborhoods over others. These civic algorithms are opaque, inaccessible, and are rarely if ever, developed through a public and participatory process [24].
>
> Sociologists such as Zukin et al. have demonstrated how geographically coded Yelp reviews reinforce prejudice against neighborhoods with people of color thereby influencing civic investments and contributing to processes of urban change such as gentrification. The algorithmic moderation of reviews is not public, and users are unaware of their role in affecting capital flows [25].
>
> Loukissas argues that filter bubbles created by Zillow reinforce existing spatial power structures along the axes of race and class. Even as the goal of the site is to give users more control over their home-buying experience, it leaves users unaware of the implications of their own filtering decisions [23].
>
> In my past work, I illustrate how safe walking apps, such as Safetipin, risk segregating neighborhoods by attempting to advance the safety of one social group while marginalizing another. Once again, knowledge about the failings of social structures and policy normalized by the app is inaccessible to citizens, as well as the app creators [10].

Without careful consideration of the spatial workings and effects of predictive systems, they risk reproducing or even amplifying historical systems of discrimination. Currently, civic algorithms are deployed in ways that prevent citizens from getting access to them or learning about their existence and effects. This limits the citizens' ability to contest, oversee, assess, or protest automated decisions that affect their lives in fundamental ways. To address concerns of algorithmic opaqueness and lack of understanding, activists [26], civic organizations [27, 28], as well as several governmental agencies [29] are calling for advancements in algorithmic transparency, explainability, and impact assessment. Their goal is to drive visibility into the design of algorithms, opening them up for critique and evaluation by both experts and everyday users.

1.1 Existing XAI and AI Transparency Approaches

Early research on Explainable AI focused majorly on technical explainablity and transparency. For black-boxed algorithms, post-hoc explanations where another human-interpretable model imitates the practices of a complex model are being designed to understand how an algorithm makes its predictions. Guidotti et al. [30] categorize post-hoc explanations into these categories: (1) Global Model Explanations where a simpler interpretable model is trained on the same data to approximate the working of the primary more complex model, (2) Outcome Explanations where the focus is on explaining one outcome or instance of prediction by providing the weight of features that contributed to it [31] or examples of inputs that would lead to a similar prediction, and (3) Counterfactual Explanations where the goal is to identify what should be changed in the inputs to get a different prediction [32]. Beyond making models interpretable, XAI efforts attempt to make known other aspects of the machine learning model such as datasets, training algorithms, source code, and performance metrics [33]. Several model [34] and data documentation [35, 36, 37, 38] frameworks, have been designed to explain AI to experts [39]. Frameworks such as 'CrowdWorkSheets' support standardized documentation of decisions when annotating datasets in a crowdsourced

manner [40]. 'Data Cards' provide summaries of datasets for various stakeholders [41]. To promote transparency of models, existing work proposes tools such as 'Model Cards' that aim to provide details about a model related to its working, use cases, and evaluation [34] and 'Factsheets' that provide an overview of facts about specific models across the AI lifecycle [42]. There is also growing work in designing open-source toolkits to identify and assess algorithmic harms and biases [43, 44, 45]. AI Fairness 360 (AIF360) [43] and Fairlearn [45] are tools that aim to help practitioners understand 'bias' metrics and allow them to detect algorithmic biases. These tools also help mitigate said biases by providing a variety of mitigation algorithms. Another tool called "What-If" uses visualizations to help users and practitioners investigate how a model will perform in hypothetical scenarios created by changes in data points [44].

More recently, the XAI community has presented the need to make AI processes visible to the general public in order to build trust in the artificially intelligent systems that guide their lives [46]. There has been a growth in work that focuses on developing methods and frameworks for user-centered algorithmic explanations [47, 48, 33]. Human cognitive abilities [49], users' explanatory needs [50, 51], users' situated real-world experiences [52], users' ability to collaborate and form counter publics [53] have been some of the guiding factors in advancing user-centered XAI research.

Such efforts have built on diverse theories and methods: interactive explainability methods [54], such as Interactive Model Cards [55], have been proposed, theories describing human cognition patterns have been employed [49], and example-based methods where data samples that informed a prediction are shown to users have been presented as helpful [56]. Additionally, several toolkits have been designed to promote user understanding of AI systems such as: TILT (transparency information language and toolkit) that organizes the information that transparency policies demand in structured ways for machines to read and users to consume [57], or AIX360 toolkit [58] that present visualizations and explanation algorithms respectively to help users understand how predictive systems work. Investiga-

tions into how XAI can support user assessment [59] and decision-making capabilities [60] are also being conducted.

These efforts have made monumental progress in the field of Explainable AI. Yet, there remain limits. First, they are not pluralistic in nature [61, 62] and do not consider the values, surrounding social systems, existing knowledges, and beliefs [63] of users in the design of explanations. Second, they may focus merely on technical transparency, disregarding the systemic factors surrounding any algorithmic decision [64]. Third, they tend to be one-time explanations that portray users as passive individual consumers [65]. And lastly, the goal of most explanations is to increase user trust rather than to promote critical thinking or action [66]. This leaves little room for democratically evaluating how the complex world we live in is simplified to be represented in the design of algorithms— who is included, who is excluded, and how cities are quantified and aggregated for computation. I discuss these limits in more detail in chapter 2.

This dissertation aims to address these limits and theorize the design of effective public explanations of civic AI tools. I want to note that in using the term 'explanations', I am not referring to the technical explanations provided through the use of algorithmic interpretability methods. Rather, I am attempting to expand our community's understanding of 'explanations' such that it is not limited to the technical understanding of the systems but is explaining the social, political, and economic development, use, and effects of civic predictive systems.

I ground my work in public safety AI systems. Specifically, I employ place-based predictive policing as a case study for this research.

1.2 Public Safety Algorithms and Harms

My research for this dissertation started with a critical analysis of a renowned safe walking app, primarily deployed in India, called 'Safetipin' [67]. Safetipin recommends 'safe' paths to users from an origin to a destination by calculating 'safety scores' for various

paths. These safety scores are calculated by aggregating crowdsourced 'safety data' such as the amount of lighting, or presence of security officers, in various locations in a city. To study this tool, (1) I draw on feminist criticisms of safety technologies, defined broadly, to identify the main issues in their framing and efficacy, and (2) I build upon initial interviews with users and makers of Safetipin to examine how the app addresses, or fails to address, the criticisms received by other safety technologies.

This critical analysis demonstrates Safetipin's capacity to (1) restrict women's movement to computationally calculated 'safe' neighborhoods and (2) reinforce caste and religion-based segregation in India [10]. By disregarding the prejudice about vulnerable neighborhoods that governs the 'feeling of safety' of its users who contribute to the crowdsourced data, Safetipin fails to situate itself in the broader historical politics of safety in the city [68] that continue to marginalize people of lower socioeconomic status and minority religions. Nonetheless, the app and its underlying information infrastructures that promote segregation, have been enthusiastically accepted and celebrated [69, 70]. This work presents a dire need to identify and understand the impact of spatially distributed data inputs and aggregations on the city and its people. Ultimately, it motivates the need to explain how emerging civic geospatial technologies, especially in the realm of public safety, organize cities and their impact on spatial segregation and discrimination.

Amongst other tools for public safety exist a variety of ML algorithms that have been developed with the hope to mitigate crime and advance citizen safety in smart cities. Popular examples include— COMPAS [71], which predicts recidivism risk for an individual; Predpol (now Geolitica) [72], which predicts geographic areas where crime is most likely to happen; Arnold Public Safety Assessment [73], which provides judges with sentencing recommendations. These algorithms, even as they aim to promote public safety in cities, tend to reinforce discrimination along the axes of race and class. Jefferson demonstrates how Predpol legitimizes the bias embedded in official crime datasets and has resulted in the over-policing of already heavily surveilled neighborhoods [74]. Risk assessment tools

build upon and reinforce the racist policies and infrastructures underlying carceral systems in the US. Additionally, they define 'risk' at the level of an individual, disregarding how 'risk' is a reflection of societal prejudice against various social groups [75]. In India, centralized systems and norms along with the subjectivities of individual police officers lead to historical, representational, and measurement bias in recorded crime data for the Crime Mapping Analytics and Predictive System (CMAPS) predictive policing tool. Further, the opaque design of CMAPS allows for discrimination against immigrant colonies and minority settlements by promoting the belief that crime rises in specific neighborhoods by virtue of the above-mentioned communities living there [76]. Transparency is a much-needed feature for the effective assessment and development of public safety algorithms [77]. Given the limited potential of techniques designed to "de-bias" public safety algorithms, there is an urgent need to make the data assemblages [78] surrounding public safety algorithms transparent and accessible to city residents and governmental bodies [76].

This dissertation aims to study the concept of effective public explanations, by specifically focusing on place-based predictive policing as a case study. Place-based predictive policing is a method that aims to support the efficient distribution of police resources in a city. Typically, the method utilizes historic crime data such as arrest reports or calls for service requests integrated with other social and environmental data to predict the location and time of future crimes. A popular example of such a system is called Geolitica (previously Prepdol) [79]. In the past decade, academics and activists have heavily scrutinized the use of predictive policing [80]. The tool has been shown to reproduce existing geographic and social biases embedded in historically discriminatory crime data [74, 81]. A recent study conducted by the Markup has found the accuracy of this tool to be less than half a percent [72]. Many others have discussed the incompleteness of data [82], the inability to validate results [83], and the proliferation of positive feedback loops [80]. Such thorough investigation of this predictive tool makes it an ideal case to ground our discussions of effective public explanations.

1.3 Research Questions and Chapter Outline

In this dissertation, I ask: *How can the XAI community support public understanding of the spatial workings and effects of civic predictive systems (RQ)?* To investigate this question, I primarily employ community-centered qualitative and participatory methods. Additionally, I build on and am in conversation with scholarship from fields such as human-computer interaction (HCI) that investigates how users perceive, relate to, and interact with AI systems to seek AI explanations; science and technology studies (STS) that shines light on the socio-technical environments surrounding AI, publics affected by AI, as well as the processes of explaining and knowing; and AI Transparency and Explainability (XAI) that has designed numerous methods, frameworks, and tools to explain AI systems to a variety of audiences.

Chapter 2 begins my investigation of the primary research question guiding this dissertation (RQ) by asking: *What qualities underlie effective public explanations of civic predictive systems (RQ1)?* My inquiry is driven by a semi-structured interview study along with an extensive review of existing scholarship on public explanations for civic predictive systems. I interview 23 participants including academics, AI activists, journalists, community and neighborhood leaders, and civic society organizations who think, write, or act on issues of social justice and AI safety, to identify the qualities and characteristics underlying meaningful public explanations of civic predictive systems. Drawing on my findings, I introduce the concept of *'good enough explanations'*. 'Good enough explanations' as understood by our participants, (1) are *situated* in the lives of diverse publics, (2) explain the complex and entangled socio-technical *systems* that predictive tools interact with, (3) involve *continuous and partial* processes, and lastly (4) empower publics to *act* in ways that promote democratic deployment and regulation of predictive tools. Such explanations, I argue, may not be complete or objective but are good enough to support publics in critically engaging with the workings and effects of civic predictive systems in service of their

goals. The rest of the dissertation attempts to understand and explore these qualities.

Chapter 3 documents my approach to study, as well as create, good enough explanations. *Form* has been considered an essential dimension of AI Explanations [84]. As such, I take inspiration from the visual history of city representation to demonstrate how *mapping* may allow us to explain geo-spatial AI systems in ways that are accessible, culturally reflexive, situated, and provide visibility into how algorithms represent cities. I employ a 'research through design' methodology and conduct participatory mapping workshops with diverse publics such as— police reform groups, city planners, neighborhood and community leaders, civic development agency, and educators— to investigate how can we, as XAI researchers, design good enough explanations. The workshops encouraged participants to question, understand, and explain place-based predictive policing on their own terms by using maps to ground discussions in the spaces they live and work in. The explanation contexts that emerged were analyzed using inductive coding, situational analysis, memo-ing, and thematic analysis. These workshops support the findings and arguments developed in chapters 4, 5, and 6.

Chapter 4 studies the 'situated' nature of public explanations and asks: *How are explanations situated in the lives of diverse publics and what does that mean for the design of public explanations (RQ2)?* Drawing on five participatory mapping workshops with diverse publics, this chapter reflects on the explanation contexts that emerge when publics question or understand AI systems. I find that when attempting to understand AI systems, publics draw on their relation with not just the predictive technology, but also the prediction domain, prediction subject, and prediction backdrops. Situated explanation needs of publics emerge out of these entangled relations. I argue that these broader relations play a significant role in supporting public understanding of civic AI systems and urge XAI researchers to consider these relations as they attempt to design systems for public understanding of AI.

Chapter 5 studies the 'systemic' nature of public explanations and asks: *How can the*

XAI community help explain the complex socio-technical systems that civic AI tools engage with? Underlying AI systems are socio-technical assemblages of materials, relations, cultures, institutions, and histories. In this chapter, I analyse the participatory mapping workshops using situational analysis to demonstrate the ability of local publics to partially explain the environments AI tools are deployed in, the cultures and norms they invade, and the lived experiences of the problems they attempt to address. Drawing on these findings, I argue that good enough explanations need not be designed in isolation *for* publics *by* XAI researchers. Instead, XAI researchers would benefit from co-creating systemic explanations *with* diverse local publics through slow and long-lasting engagements.

Chapter 6 reflects on the methods I use to develop good enough explanations to reconceptualize explanations as explain*ing*— a continuous and partial process that supports the development of *situated* understandings of AI *systems*. I provide a detailed account of three design dimensions I believe should be considered as XAI researchers design spaces and systems that mediate processes of explaining: (1) designing explaining sites (2) designing explaining media, and (3) designing explaining interactions. I briefly discuss how explain*ing* may influence public 'action'. I end by providing a list of questions that XAI researchers can consider as they attempt to create good enough explanations for civic predictive systems.

Lastly, I conclude by providing a summary of the research conducted in this dissertation, highlighting limits of this work, proposing relevant future work, and positioning the work in relation to my background and motivations.

Together, these chapters attempt to make three primary contributions to the field of XAI: (1) theoretical contribution by conceptualizing what *good enough explanations* mean for public understanding of civic predictive systems, (2) empirical contribution by reporting how diverse publics may understand, question, and explain civic predictive systems, and (3) methodological contribution by demonstrating and recommending strategies for co-creating a good enough understanding of civic predictive systems.

My hope is that this work will support XAI researchers, policymakers, and community leaders in designing systems and spaces for public engagement with civic predictive systems. Such engagements, as stated in Chapter 2, should work to promote democratic public action toward contesting, redesigning, regulating, or discontinuing predictive systems that may cause societal harm in fundamental yet currently invisible ways.

CHAPTER 2
GOOD ENOUGH EXPLANATIONS

I begin this dissertation by asking: What qualities underlie effective public explanations of civic predictive systems (RQ1)? In this chapter, I report on interviews with people who me and my collaborators see as stakeholders in civic AI, such as academics, journalists and leaders in civic organizations and neighborhood associations, all of whom seek pragmatic explanations for the predictive technologies that are poised to change the communities in which they work. These interviews shed light on the qualities and concepts that such stakeholders are looking for in effective explanations of civic predictive systems. As mentioned in Chapter 1, this research draws on place-based predictive policing as a case study. Participants in this study identify the following four questions as imperative for creating effective public explanations: (1) who receives and creates an explanation; (2) how is an explanation developed and shared; (3) what does an explanation explain; and (4) what are the goals and impacts of an explanation. In following, I articulate the need for situated, systemic, continuous and partial, and actionable public explanations of AI for the civic realm. I call these 'good enough explanations'.

2.1 Introduction

As predictive systems increasingly inform civic decision making [4, 8, 7], activists, academics, as well as policymakers are calling for public involvement in their assessment and regulation [85, 86, 87]. However, a limited understanding of the existence, workings, and impacts of predictive systems has made it challenging for citizens to democratically participate in the decision-making surrounding the use of these tools [86]. As described in Chapter 1, Artificial Intelligence (AI) explainability, transparency, and interpretability methods have been proposed to promote public understanding of AI systems. However,

recent scholarship has demonstrated the limits of existing approaches for meaningfully supporting public knowledge, usage, and assessment of civic AI systems [65, 86]. They discuss how current methods, while influential, (1) design for individual users as static and universal consumers of AI explanations disregarding their pluralistic and situated perspectives, (2) aim to merely provide technical explanations of AI separate from the broader socio-political systems that surround it [64, 88], (3) are one-time explanations that may be overwhelmingly lengthy or convoluted, and (4) aim to develop user trust rather than problematize trust and develop critical thinking or promote action [89, 90]. These limits highlight the need for the Explainable AI (XAI) community to center local publics in the study and design of effective explanations.

To address this epistemic challenge, this chapter asks: *What qualities underlie effective public explanations of civic predictive systems (RQ1)?* In this paper, I report on twenty-three interviews with people who I, along with my collaborators, see as stakeholders in civic AI, such as academics, journalists, and leaders in civic organizations and neighborhood associations, all of whom seek pragmatic explanations for the predictive technologies that are poised to change the communities in which they work. As I have mentioned, this dissertation focuses specifically on civic predictive systems, i.e., predictive systems used for civic purposes, because of their omnipresent yet concealed nature. The lack of direct interaction with such technologies makes democratic oversight and assessment especially difficult, thereby presenting an urgent need for investigation. Specifically, I focus on place-based predictive policing and its impacts, that have been studied at length making it a rich case to help build our understanding of effective explanations [91, 92, 93].

Participants in this study identify four questions as imperative for creating effective public explanations: (1) who receives and creates an explanation; (2) how is an explanation developed and shared; (3) what does an explanation explain; and (4) what are the goals and impacts of an explanation. This chapter, including my report of current XAI limits in the following section, is structured around these four questions.

I also identify three challenges faced by the broader XAI community in designing effective explanations: (1) limited access to information about civic predictive tools, (2) lack of awareness, interest, or availability amongst the public to engage with predictive tools, and (3) lack of consensus on best ways to regulate tools and mitigate their harms.

Through an interpretive analysis of the findings, I conceptualize effective public explanations as processes (1) that are *situated* in the lives of diverse publics, (2) explain the complex and entangled socio-technical *systems* that predictive tools interact with, (3) involve *ongoing and partial* processes, and lastly (4) empower publics to *act* in ways that promote democratic deployment and regulation of predictive tools. I bring these dimensions together by introducing 'good enough explanations' as a tool that embodies these values. I borrow the concept of 'good enough' from the work of Donald Winnicott, a pediatrician and psychoanalyst, who coined the term 'good enough parenting' [94]. He contrasts a good enough parent with a perfect parent and notes that a good enough parent does not meet every need of their child. This, he argues, helps their child be more resilient. I want to highlight that Winnicott conceptualized the term 'good enough' in the 1950s, a time before the second wave of feminism, when discriminatory gender roles were condoned by the society. His work, then, characterized women as the sole care taker of an infant in a nuclear family [95]. What I take from him work, however, is his semantic conceptualization of 'good enough'. The term 'just good enough' was also recently used by Gabrys et al. [96] who introduced 'just good enough data' to describe citizen data created via limited means and methods that may offer new ways to relate to data and mobilize them. Since algorithmic explanations too cannot be perfect [88], this chapter presents good enough explanations as explanations that are not 'complete' or 'universal', since that remains impossible, but are good enough for diverse publics to consciously engage with civic AI systems. My hope is that such explanations can help overcome the epistemic barriers presented by opaque algorithms [97]. Ultimately, I define good enough explanations as *ongoing processes* that allow *diverse publics* to *partially* engage with *features of predictive*

systems and *assess such systems in relation to their communities*.

In what follows, I provide a brief background on the limits of current Explainable AI (XAI) and AI transparency work, that motivates this chapter. Next, I describe the methods that this chapter employs, followed by the findings. The findings section also includes a description of the challenges presented in developing effective explanations. Lastly, I end by analyzing the findings to propose and define the concept of good enough explanations.

2.2 Background

Existing transparency and explainability methods have made noteworthy progress in identifying and communicating the workings of and harms associated with predictive systems. I provided a detailed account of such methods in Chapter 1. However, more recently, scholars have questioned their effectiveness. There is limited work focusing specifically on public explanations of civic systems through the lens of stakeholder-centered qualitative methods.

In this section, I present an overview of the limits of approaches aimed at promoting the explainability and transparency of algorithmic systems as discussed by the broader Responsible AI literature. This is followed by a summary of few but growing works that focus specifically on public transparency and participation. The scholarship discussed is not an exhaustive, but a representative sample of works relevant to this research.

2.2.1 Limits of Current XAI Approaches

Who is the explanation for?

Early research on explainability focused on designing explanations for internal stakeholders who perform model debugging and improvement [98]. More recently, there has been an increase in efforts to widen the scope of XAI and design user-centered explanations [47, 99]. However, despite their widespread and note-worthy contributions, these approaches remain limited. They tend to focus on individuals as indistinguishable consumers of ex-

planations assuming similar needs and knowledges, disregarding the diversity and dynamic nature of explanation needs of various social groups [100]. The objectivity and seamlessness of current explanations discount the role of pluralism in the design of explanations [62, 101]. They consider explanations to be universal and one-size-fits-all artifacts [102]. In doing so, they fail to conceptualize explanations as social artifacts [47]. Situated considerations are not reducible to optimizing for efficient and accurate transfer of information [88].

Additionally, current efforts focus on individuals as consumers of information outside of a social group, community [86] or the broader public. Focus on individuals prevents social groups from taking collective action and allows AI developers to escape the consequences of any misconduct [103].

What is being explained?

For black-box algorithms, post-hoc explanations are being designed where another human-interpretable model imitates a more complex model, in order to understand how an algorithm makes its predictions [30]. The majority of these efforts focus on making transparent the technical aspects of AI systems [65]. Beyond making models interpretable, they attempt to make known other technical aspects such as datasets, training algorithms, source code, and performance metrics [99]. However, merely technical transparency has been considered insufficient to support effective decision-making as it disregards the contextual factors surrounding any algorithmic decision [64]. Even when technical transparency is provided, end users may not be equipped to connect these materials to the effects of the AI system on their communities [64, 104]. Another aspect of transparency, discussed by Eypert and Lopez, is its scope. Transparency efforts do not consider elements of an AI system that are needed for democratic oversight. For instance, they do not account for media discourse surrounding the use of an AI system that may create harmful or incorrect expectations [86].

How is an explanation shared?

In this work, I do not focus on interpretability techniques such as post-hoc explanations. Instead, the focus is on approaches to engage users in explanations. Current explanation methods tend to be passive and isolated [65]. They only allow for a uni-directional flow of information from the makers of technology to the consumers of technology, experts to non-experts, engineers to users [65, 86]. Additionally, they tend to be overly detailed and technically comprehensive. However, explaining an AI system in its entirety and merely making visible the 'black box' does not guarantee that a user understands the system they see [105]. An overload of information about predictive tools offered quickly can not only be unnecessary but can overwhelm citizens [106]. At other times, despite the system's goal to be transparent, the information may be difficult for users to find or the information may not be relevant, clear, or consistent [4].

Why do we need explanations?

One of the primary goals of user-centered explainability has been to promote trust in predictive systems [107, 108, 109]. Cultivating trust in users has been considered useful for several reasons such as increasing adoption, taking advantage of the full range of AI's potential including an increase in productivity, or is considered intrinsically valuable [110]. However, increased transparency can result in users over-trusting predictive systems and placing more value in a prediction than it deserves [111]. XAI efforts may also mislead users into thinking that they have more control over predictive systems than they actually do [112], thereby shielding developers and the algorithms from any blame [113]. More fundamentally, scholars also remind us that trust is only desirable if it is deserved [114]. 'Warranted distrust' can be more productive when we engage with imperfect AI. It can help caution users about the limited capabilities of predictive systems [90]. Recent scholarship has highlighted the importance of problematizing public trust [89] and developing critical thinking [66].

The limits of XAI work summarized in this section serve as my motivation to ask: How can we address these XAI limits and design effective public explanations? Below, I present related work that provides us a foundation of research to build upon.

2.2.2 Public Engagement With Civic AI Systems

Current XAI scholarship, including user-centered XAI, has given little attention to public-centered explanations of civic AI. However, slowly but increasingly, scholars are coming up with toolkits, frameworks, and guidelines to meaningfully engage the public in civic AI. The Algorithmic Equity Toolkit (AEKit) allows the public to define AI, determine if a tool qualifies as an AI, learn about AI's constituent parts, and develop a vocabulary of probes to ask critical questions about AI [115]. The Model Card Authoring toolkit supports community members in collectively deliberating over the design of models, the opportunities they present and their limitations, and how can they serve the needs of their specific communities [116]. WeBuildAI supports communities in coming together in a participatory manner to define AI policy [117].

Frameworks and guidelines for public engagement have also been proposed: Michele Gilman published a report with Data and Society that identifies eight principles for meaningful public participation in AI governance which includes methodological and motivational components [118]. A similar report published by the Data Justice Lab discusses how the construction of mini-publics, citizens assemblies, citizens juries, and participatory budgeting, among other community groups, can support civic participation in relation to algorithmic decision-making [119]. Anna Colom of the Ada Lovelace Institute also provides starting points for meaningful public participation such as by designing systemic processes for continuous public participation in AI decision-making [120].

This study builds on the scholarship above in the following ways: (1) it focuses explicitly on explaining *civic* AI systems, (2) it directly engages with diverse stakeholders to conceptualize our understanding of effective civic AI explanations, and (3) it organizes

our qualitative findings and existing literature to term 'good enough explanations' that can serve as a useful tool to inform the design of public explanations.

2.3 Methods

This study draws on twenty-three semi-structured interviews with people from academia, civic organizations, activist communities, non-profits, journalism, neighborhood associations, etc. (described in detail in Table 2.1). All participants that were interviewed have written, thought or acted extensively on issues of predictive systems, criminal justice, social justice, and city life. They were chosen due to their presence and experience as members or leaders of communities with stakes in the harms caused by civic predictive tools. They were identified through a broad search for stakeholders that could speak to these issues on a local or national level in the United States. They were recruited through direct contact and referrals.

The interviews lasted a total of thirty to ninety minutes. The study was performed in Atlanta, GA in the United States. These interviews included, but were not limited to, the following discussion prompts:

- What, if any, is the role of citizen-centered AI transparency in the design and deployment of just civic predictive tools? What is at stake and why? What are the challenges? Why?
- What is needed to promote meaningful citizen-centered transparency? Is partial understanding enough? What does democratic control over AI look like?
- Do you or your association think about AI use by cities? Why? Why not? Do you think there is a need to do that?
- What do you know about the use of AI for public safety in your neighborhood? What questions do you have? What forums exist to offer information on the use of AI by cities?

Table 2.1: Social roles and expertise of the stakeholders we interviewed in our study

	Profession	Expertise
P1	Philosopher, Academic	Studies AI ethics and predictive policing
P2	Philosopher, Academic	Studies technology ethics, policy, and governance
P3	Philosopher, Academic	Studies policing and justice
P4	Sociologist, Academic	Studies impacts of policing and surveillance
P5	Journalist, investigative researcher, digital rights non-profit group	Studies law enforcement technology and government transparency
P6	Lawyer, member of an academic non-profit that partners with community to promote public safety	Designs systems for transparency and democracy in policing
P7	Data scientist, non-profit, focused on human rights violation	Works with community partners and grassroot orgs to advance human rights
P8	Leads a non-profit in [anonymous state] focused on digital rights	Develops frameworks for evaluating new surveillance and carceral technology proposals
P9	Case worker, Innocence Project	Helps free innocent people sentenced to life in prison in southern United States
P10	Communications and Technology Studies, Academic	Studies socio-economic inequalities in urban techn
P11	Director, [anonymous city]-based non-profit organization	Works to address issues of transparency, police accountability and digital security
P12	Research Director, non-profit based in [anonymous state]	Litigates and advocates against technologies impacting marginalized communities.
P13	Ex- security chair and Vice President, Neighborhood Association, [anonymous city]	Leads a citizen advisory council, contributes to neighborhood well-being.
P14	Head of the Public Safety Committee in an [anonymous city] neighborhood	Leads a citizen advisory council, contributes to neighborhood well-being and safety.
P15	Member, Neighborhood Association in [anonymous city]	Discusses issues of significance for their neighborhood as part of an association.
P16	Board member, Neighborhood Association in [anonymous city]	Participates in decision making related to neighborhood's priorities and well-being.
P17	President, Neighborhood Association in [anonymous city]	Crafts and leads civic initiatives related to neighborhood needs
P18	NPU Chairperson in [anonymous city]	Leads civic initiatives related to neighborhood needs
P19	Director of Civic Engagement, Non-profit	Empowering and elevating African American communities in [anonymous city]
P20	Director, Non-profit, [anonymous state]	Empowering individuals and revitalizing communities in [anonymous state]
P21	Director, Non-profit	Works to reduce arrest and incarceration of marginalized people in [anonymous city]
P22	President, Civic non-profit	Partners with under-resourced neighborhoods in [anonymous city] to support community development
P23	Manager of a sub-group in a non-profit	Advances policies and actions that promote racial equity in [anonymous city]

The interviews were either (1) conducted over Zoom and transcriptions of the interviews were recorded [P1-P12, P18-P23], (2) conducted over the phone with active notetaking [P13-P16], or (3) conducted in person with active notetaking [P17]. The decision to not record audio or video during the interviews was taken after a pilot interview where I found that recording the interview was making participants uncomfortable or hesitant in their responses.

2.3.1 Data Analysis

Following the thematic analysis method proposed by Braun and Clarke, I first familiarized myself with the transcriptions and notes. Next, I coded the data using an inductive method identifying both semantic and latent codes such as: 'explanation needs vary with stakeholders, 'explanation should promote local expertise', 'explanations should provide information about procurement', 'explanations should not be too technical', 'explanations are necessary but not sufficient', and more [121]. These codes were then organized into themes based on the characteristic of the explanation they referred to– who receives and creates an explanation?; what does an explanation explain?; how is an explanation developed and shared?; and what are the goals and impacts of an explanation? These themes inform the structure of this chapter. I also describe the challenges in designing effective explanations.

Next, I identified relationships within these themes through the process of memoing [122]. This process revealed patterns that described the qualities inherent to effective explanations. I term explanations embodying these qualities 'Good Enough Explanations'. I borrow the term 'good enough' from other domains where it has been demonstrated to be a useful concept to help navigate epistemic and actionable imperfections. Here, I aim to describe what 'good enough' could mean for explanations that remain imperfect [88].

2.3.2 Methodological Considerations and Limitations

This was a qualitative and interpretive research study and doesn't aim to be reproducible. I conducted the interviews and analyzed the data. I was familiar with the data and its contexts in ways that allowed rich interpretive analysis.

The interpretive and exploratory approach also motivated my selection of participants where my goal was to learn more about public explanations of civic AI from a wide range of stakeholders. I aimed to recruit participants with diverse experiences and relationships with predictive policing systems across their areas of participation and expertise. In this chapter, I do not report on demographic information about the participants, because this was an initial, exploratory study that focused on social roles emerging around predictive technologies, rather than personal experiences. In later work, I plan to explore the importance of gender, race, age, and other demographic distinctions in the creation of explanations for civic AI. This study does not include quantitative evidence and as such I do not present the perspectives of an evenly distributed sample of participants. Instead, this study serves as a starting point to investigate public explanations of civic AI systems and I propose concepts that can be useful for my fellow researchers to build on, nuance, or challenge.

2.4 Findings

In this section, I report on the following questions and concepts that the stakeholders in this study considered essential for effective explanations of civic AI systems: (1) who receives and creates an explanation?; (2) what does an explanation explain?; (3) how is an explanation developed and shared?; and (4) what are the goals and impacts of an explanation? Participants also discuss challenges faced by the research and activist community in designing effective explanations namely: (1) limited access to information about civic predictive tools, (2) lack of awareness, interest, or availability amongst the public to engage with predictive tools, and (3) lack of consensus on best ways to regulate tools and mitigate

their harms.

2.4.1 Who receives and creates an explanation?

For designing 'public' centered explanations, participants highlight the importance of considering who comprises of the 'public' [P4]. P3 talks about how those most at-risk may not be the ones involved in decision-making, and explains:

> "You'll often have people appealing to vaguely democratic justifications for police power, like the community supports this, right? ...but the people who support the bills are often not the people who will be subject to the police power." [P3]

It then becomes essential to carefully consider who the explanations are being designed for and therefore who is included in public conversation about predictive technologies. P4 reinforces this by speculating the following scenario: if we were to involve people from home-owners associations, we would self-select people who own homes and therefore are relatively richer [P4]. The choice of the 'public' made by the XAI researchers then becomes essential and consequential to the just inclusion of communities in algorithmic oversight and governance.

P2 describes how effective transparency is shaped by the needs and goals of publics in varied contexts:

> "What is needed in terms of transparency is always a function of what people are trying to accomplish...transparency needs to be molded in very specific ways so that people are being provided with particular pieces of information that are useful." [P2]

Participants discuss in detail the diversity of explanation needs that various publics may have. While some people who want to audit the tools may want to know about the source code, data, and models used [P2]; others who are affected by the tool while being unaware

of its existence may merely want to know about the presence of the tool along with its goals and effects [P13, P14, P17]. Civic organizations working to address issues of justice and inequality want to know how these tools may make their tasks harder [P20]. Police reform groups like policing diversion teams may "*want to know if these tools are going to make the police more or less likely to find themselves in situations where they are responding to diversion calls*" [P2]. Management in large police departments *"should be able to at least explain from one person to another the different elements that are being used"* [P5]. Other groups considered important stakeholders in need of explanations included: police advisory boards, neighborhood associations, city commissions, police officers, anti-police violence and anti-racial discrimination organizations, homeowners' associations, district attorneys, and public defenders.

2.4.2 What does an explanation explain?

According to participants, explanations should make known the financial, historical, political, and social aspects of an AI system. They ask the following questions about place-based predictive policing systems:

Financial

What is being spent on procuring these tools and where is the money coming from [P6, P7, P13]? How is taxpayer's money being used [P4]? What is the decision-making process that impacts the procurement of predictive tools for civic purposes [P11]? Vendor contracts signed by governmental agencies are opaque [P7] providing little transparency into how money is spent on these tools and who makes these decisions. Such gaps in knowledge, as stakeholders discuss, may prevent the public from participating in decisions about how taxpayer money is used.

Historical

What are the origins of predictive technologies [P11, P23]? What existing systems and structures gave rise to these tools? How did we get here? [P23]. Knowledge about the historical grounding of these tools can help publics assess the harm caused by their predecessors and anticipate the possible impacts of the predictive tool in question [P11]. By understanding the effects of tools that came before predictive policing, and were deployed in a similar domain, such as body cams, we can evaluate how our current approaches are similar or different [P4].

Socio-political

What are the motivations of the makers and buyers of these tools [P11, P9]? What are the priorities for the police departments [P3]? Are we policing sexual assault and homicide or sex work and decriminalized drug use [P3]? Technologists may try to find a problem that they think predictive tools can solve instead of thinking of ways to address problems that a community currently faces [P6]. It is essential to understand what community concerns are and if and how the predictive tools address said concerns. As P6 describes, predicting crime may not be a useful course of action in several cases:

> "At the end of the day, you're diagnosing a social problem and sending police as an answer to that problem when they can't really answer it... we need to find where to send extra social services, where to send employment resources, where to send, you know, other types of public resources, assistance to neighborhoods that are suffering from those problems." [P6]

Participants also raise questions about the impacts of predictive tools on diverse social groups: Who is impacted by these tools and how [P7]? What is the cost of incorrect predictions and who bears those costs [P7]? What are the civil rights and liberties cost? What are the racial justice costs [P6]? What are the discriminatory effects of using this

tool vs not using it [P7]? How do these tools interact with other parts of the criminal justice system [P3]? Or other departments like housing or child welfare [P5]? How is its effectiveness/efficacy measured or evaluated [P7, P1]? Have there been studies to prove the public safety benefits of these tools [P6]? What are the safeguards in place for when the tool has negative impacts [P3]? Participants prioritize the need to learn about the impacts of public safety tools, including the validity and reliability of predictions, over the need to uncover the black box and understand how the model works [P6, P15]. They demand studies to evaluate the predictive tools and provide concrete evidence of their effectiveness and impacts on social groups [P6].

Socio-technical

Participants discuss the need to understand the socio-technical elements that affect the working and use of predictive tools. Throughout the life cycle of the product, participants ask several questions:

Existence of tools: What systems are being used by different police departments [P3, P4]? When do the contracts end or are renewed [P13]? Basic descriptive information about what systems exist and are used by a city is essential for the public to engage with the use and regulation of the tools.

Data Type and Limits: What data is being fed into the system [P9]? Is the data specific to one's own community [P2]? How does that impact the predictions [P9]? What is over-represented and under-represented in data [P7]? Calls for service data may embody the assumptions of who and what looks suspicious to people who call 911. Arrest data are a product of targeted policing in black and brown neighborhoods. Learning about the data that a predictive system is trained on can support the public in identifying the limits of the tool and their potential impacts on their communities.

Prediction Goal: What kind of crimes is the tool predicting [P9]? P9 explains *"it matters if crimes are found by police or are community reported"*. Since drug or intoxication

crimes are officer-initiated crimes, the mere presence of police officers in a neighborhood can lead to an increase in the documentation of such crimes [P1]. This shifts the focus of police away from crimes such as assault and murder to petty crimes that are being decriminalized in several states.

Action: What are police instructed to do when they are in a crime hotspot [P1]? What does the successful use of a predictive tool look like [P5]? Are we collecting data about misidentifications, false arrests, police interactions, and pedestrian stops [P5, P6]? A predictive tool causes harm not through the act of making a prediction alone, but because of how the predictions change space and people in space. P7 discusses ShotSpotter, a tool that claims to identify gunshots and notifies police forces, *"in the case of ShotSpotter, whether there was a gunshot or not, the police show up with the intent and state of mind of responding to gun violence and this is when really awful things can happen"* [P7]. Therefore, it is essential to assess the protocols followed by police forces as they interact with predictions [P6].

To summarize, for participants, meaningful public explanations provide informatiom about (1) finances related to funding predictive tools, (2) historical insights about the origins of predictive tools, (3) socio-political details including motivations and impacts of predictive tools, and (4) socio-technical details such as prediction goal, data use, and action protocols related to a predictive system.

2.4.3 How is an explanation developed and shared?

Participants discussed the quality of methods needed for designing and delivering effective public explanations. P7 questioned the potential of one-off explanations:

> "I don't think there is a version where there's some educational thing and everybody in the community now understands this and it's like, okay, now go off and monitor your local police or whatever." [P7]

According to P7, such one-time explanations are ineffective in supporting public over-

sight over policing systems. Another feature discussed by participants relates to the extent of information explanations provide. P3 explains the importance of limited information and states:

> "I think viewing them (predictive tools) as procedures that you can assess without knowing how the nuts and bolts of everything work, that is important." [P3]

Communicating the technicalities of how an AI system works may not be possible or may be unnecessary for the purposes and goals of local publics. As P6 states: *"It (AI) can be a very intimidating language to try to enter..."* [P6]. Therefore, according to our participants, explanations should limit and present themselves in ways that overcome the barriers of the language of AI, standing as a gatekeeper, preventing citizens from overseeing and assessing AI tools.

P6 also emphasizes that not everyone needs to know all about the workings of predictive systems.

> "I don't think that there needs to be the same level of responsibility on each individual citizen to be able to sort of go toe to toe with the vendor or the or the, you know, police agency to say, you need to give me this, you need to give me that." [P6]

Diverse publics, then, may prefer different languages and focus points as they encounter explanations of AI systems.

2.4.4 What are the goals and impacts of an explanation?

Participants emphasize that explanations are not an end in themselves. Instead, they serve as starting points, as stated by P4: *"I think there is a little bit of like false promise of transparency.. you absolutely have to have some of that in order to even start but that.. it's sort of like the starting point rather than the final product"* [P4]. Explanations serve

as precursors to other goals. For instance, a popular goal of transparency is to allow the public to hold tech makers and government officials accountable [P6]. But accountability doesn't simply flow from transparency [P4]. Mechanisms that support accountability need to be purposefully designed around transparency efforts.

According to participants in this study, effective transparency via public explanations can serve many goals:

Dialogue

Explanations can support productive and meaningful discussions amongst various stakeholders and help people share their experiences in relation to the tools [P7]. P7 stresses the need to be involved in discussions with community groups such as those most impacted by these tools alongside policymakers who need to be informed and knowledgeable enough to represent the needs of the communities they serve [P7].

Reflection and empowerment

People believe that predictive systems are more accurate than they actually are [P10]. Explanations can make these tools appear less magical [P11, P12] and support people in thinking of them merely as police allocation tools that may carry the same problems as traditional police allocation efforts. The explanations can help people reflect on what the tools are over-promising and how the tool impacts the practice of policing. Does it improve trust in police and promote community policing [P9]? A lack of understanding, P5 argues, would mean that one is placing their trust in the creators of these technologies who may have profit-oriented interests [P5].

Explanations can also *"help people see their own expertise as useful"* [P7] in assessing and regulating predictive tools in relation to their lives. They can transform citizens' thinking *"I'm not an expert in technology. How could I possibly work to hold these people accountable"* [P2] to one where they see themselves as local experts who are able to effec-

tively oversee tools that impact their communities and neighborhoods. As P6 states below, such local perspectives are essential for the responsible implementation of predictive tools.

> "...there is a lot of learning to do from community members and I think that lived experience of what's working in a community and what's not is essential to good policy and can give a ton of insight." [P6]

Oversight and Assessment

Participants express the importance of public understanding in overseeing the responsible and equitable use of predictive tools. As P2 states below:

> "Without awareness about these kinds of technologies, and without some level of understanding about how they're being used, there's no possibility for democratic oversight...Political change comes from getting individual people, regular citizens, involved in their local communities, and without an understanding of the existence of these tools, of who is making decisions about how to use them, how they're operating and so on, people can't mobilize around injustices that the tools help to perpetuate." [P2]

According to P2, an effective understanding of predictive tools creates possibilities for democratic oversight. Explanations can support citizens in helping evaluate the tools that affect their lives [P7]. Further, they can help protect citizens' civil liberties [P6]. They also introduce communities to some of the critiques about these tools [P7]. Some participants state that it is not essential for every citizen or user to know how a system works but that there is a need for audits of predictive systems by various domain experts [P3, P5].

Regulation and change

Explanations can help people ask knowledgeable questions to tech vendors and policymakers [P6]. Explanations can support the public in thinking about what is needed to reduce

the harms, reap the benefits of predictive tools [P6], and organizing against algorithmic injustices [P2]. Even the process of attempting to explain is helpful because as P5 states "Sometimes the fact that there isn't an answer to the question is enough of an answer to that question to know that they shouldn't proceed" [P5]. Ultimately, explanations can help create a "place of conflict" [P11], providing citizens with the vocabulary to understand the tools that affect their lives and demand change.

2.4.5 Challenges in creating explanations of civic AI

Despite the progress made in the fields of XAI and AI transparency, the development and distribution of effective explanations remains challenging. Firstly, it is difficult access information about these tools. Secondly, many people may not be interested, available, or able to participate in discussions about the tools. And lastly, a lack of consensus on the best ways to regulate the tools can hinder collective organization efforts. I elaborate on these challenges below:

Access to Information

It is challenging to access information about predictive tools for several reasons. These tools are often developed by private organizations that are not obligated to share information [P2]. P5 elaborates on such lack of requirement to obligate:

> "It's a problem in our democracy and a problem for transparency because, as you have likely seen, one of the biggest issues, one of the most frequently cited exemptions when we are trying to understand anything about these systems tends to be the trade secrets. The reason that these algorithms aren't accessible or legible by the people using them is because the companies that are selling it don't want them to be accessible for anybody including the police departments." [P5]

Most information is protected from the public eye so as to either protect trade secrets from competitors [P5, P3] or keep people from taking advantage of transparency efforts to bypass the tool's security measures. Such opacity, justified by the need for security, not only prevents the public from holding police accountable but also reduces public trust and confidence in the institution of policing [123]. Despite such restrictions, activists, journalists, and academics attempt to access the limited information available through Freedom of Information Act (FOIA) requests. Unfortunately, they remain unsuccessful in many cases as they either lack the legal resources to successfully submit requests and seek information, and/or the police departments fail to cooperate in good faith or respond in reasonable time [P3]. There also exists little incentive for police departments to share information about the use of these tools. P10 explains their lack of cooperation by saying *"the less they share, the more protected they are"*. Even if one can identify some information about these systems, the information quickly becomes outdated [P4]. Such roadblocks can prevent people from continuing to take time out in pursuit of transparency [P5].

Ability to Learn

Not everyone is aware, available, interested, or able to learn more about the use of predictive tools in their cities. It may not be evident how civic predictive tools affect everyday citizens. And thus, citizens may not understand the need to think critically about these systems [P4]. As such, the development of meaningful explanations needs to be complemented by efforts to help people understand how these tools affect their lives [P4]. Several community experts confess that they would like to spend more time thinking about AI and its effects and are therefore interested in participating in events that provide them access to explanations [P20, P22, P23].

Whether or not groups can learn about predictive tools and take necessary actions collectively also depends on if they have the social capital to organize themselves in ways that make them approachable [P3]. While there are XAI efforts that focus on policymakers

or workers who directly interact with technologies, there is an alarming scarcity of work that engages with communities that are not organized in an official capacity simply due to logistical difficulties. Additionally, such tasks can just be overly demanding [P3]. People may also simply not be able to take out time to learn about predictive tools and participate in necessary actions. P8 explains: *"If you're asking for just general volunteers it can be hard to find people that have time on their schedule, they really have to have a reason to be there...You have to find volunteers who are knowledgeable and committed and really are willing to keep the city honest"* [P8].

However, not everyone has the privilege of choosing not to participate. Those who are active and committed to the processes are generally the ones who are most affected by these tools and have the greatest to lose [P8]. This also implies that the burden of responsible deployment of technologies is put on communities of color and high poverty who have already had their trust destroyed by the police [P6].

Defining Action

An essential component of just good enough explanations is their role in promoting action. However, it is challenging to act in unison when communities who are critical of predictive tools are unable to reach a consensus on the best ways to regulate these tools. P4 discusses at length the diversity of opinions and perspectives she came across during her empirical work with police departments, police reform groups, police abolitionists, and other policing-relevant social groups. She says *"Different publics have different conceptions of what is public safety and what constitutes good use of taxpayer money"* [P4]. She elaborates on her comment with an example of conflicting perspectives and adds:

> "Abolitionist orgs would be like, we don't care about regulations and policies and like these bureaucratic things that only serve to further institutionalize the state surveillance ..whereas more reformist groups would say, what are the mechanisms for public comments and, how can we build in accountability on

> the front end and how can we do audits or ongoing evaluations or be able to even vote in an educated way by having some sort of algorithmic impact assessment on the front end. So that's where it becomes kind of difficult because it's like...yeah, there's all different kinds of things that you can do, but not everybody agrees with like what the right route is." [P4]

As P4 demonstrates, different social groups may not agree on what are the best ways to design, develop, and regulate predictive systems. Additionally, people who may be most affected by these tools may not even want to engage in such discussions. P22 explains that the communities he serves do not trust the police enough to get involved in the processes of improving or assessing policing tools. Their perspectives, then, may never be considered as we develop predictive tools. Therefore, even with good explanations and transparency, the public as a whole may not be able to collectively organize and demand change.

2.5 Good Enough Explanations

In this section, I interpret the findings of this study, alongside relevant scholarly literature, to identify qualities and concepts underlying effective explanations of civic AI. I term such explanations 'good enough explanations' as they may not be 'complete' or 'universal', but are good enough for the public to consciously and critically engage with civic AI. I draw on real-world public explanations of predictive tools for policing to exemplify how these qualities may play out. This conceptualization of good enough explanations aims to be generative and these examples only serve as starting points for further developing these concepts.

2.5.1 Good Enough Explanations are Situated

Stakeholders in this study describe how explanation needs differ based on who consumes an explanation and in what contexts [124]. Focused on the realm of civic AI, they add nuance to current works that challenge the universal nature of explanations [102]. User-centered

XAI scholarship has also called for the design of explanations based on user expertise, personas, beliefs [47], existing knowledges [125], and goals [105]. Additionally, local publics themselves offer unique and local explanations of how civic AI systems may work [126]. Their role can be extended from receivers of explanations to co-creators.

As such, this study understands good enough explanations to be situated in the lives, experiences, and settings of diverse publics. Good enough explanations are good enough for *someone* and thus are defined in relation to their contexts. Determining who an explanation should choose to center is then an important, yet often neglected, choice [125]. I highlight the need to especially center the needs and knowledges of communities most at risk of being affected by civic AI.

In Oct 2023, an article by Aaron Sankin and Surya Muttu was co-published by the Mark and the Wired [127]. It describes their examination of 23,631 Geolitica predictions and calculates a success rate of less than half a percent. Here, I see two examples of good enough explanations: one provided *to* the writers of the article and the other provided *by* them. The writers discuss how they had limited transparency as only two out of thirty-eight police departments, that they reached out to, provided patrolling data, that ultimately informed their analysis. Yet, the partial explanations allowed them to launch an inquiry into the efficacy of the tool for specific cities. On the other hand, their analysis of the tool can now serve as a good enough explanation for police departments looking to make a decision about the potential adoption of predictive policing.

2.5.2 Systemic

When asked what participants in this study think diverse publics may want to know about civic AI, they brought up several financial, historical, political, and social aspects of an AI system. Fellow scholars have called for the design of explanations that go beyond the technical aspects of the AI system [128] and focus on explaining socio-technical elements of AI [110]. They have highlighted the need for an explanation of 'milieus', the environment,

where the human-AI interaction takes place [129].

As such, this study understands good enough explanations to be *systemic*. They do not merely aim to explain the technical aspects of a predictive tool. Rather, they make known a more systemic view of the tool including its financial, historical, political, and social aspects.

Baykurt et al. compare New York City's (NYC) and Seattle's civic efforts that focus on algorithmic transparency and algorithmic impact assessment respectively [130]. They highlight the problems NYC ran into and suggest that since Seattle's efforts were not focused on opening up the black box of predictive systems, they were able to better investigate the community-based harms associated with predictive tools. Yet, they consider explanations good enough when they can consider other systemic factors such as the social roles and power relations within the agencies that deploy these tools. This may help overcome the need to open up opaque AI systems. Several other public explanations of AI tools for policing highlight systemic assemblages surrounding predictive models, explaining the origins, financial costs, and effects of the tools [131].

2.5.3 Ongoing and Partial

When attempting to understand civic AI, stakeholders in this study seek parts of an explanation delivered over extended periods. Meaningful transparency has been suggested to be a "never-ending endeavor" [86]. Understanding of an AI system develops over time—"through repeated exposure and interactions" [129]. It may last beyond the product's lifecycle throughout people's lives in the form of AI literacy efforts [132]. Such slowness, over a longer period of time, invites reflection, develops critical understanding, and supports the public in actively thinking about their interactions with technological artifacts [133], while reducing over-reliance and trust in automated predictions [134, 135]. Additionally, stakeholders suggested the need for limited, but relevant, explanations. As Miller describes, people tend to select a few primary causes as an explanation rather than desire a *complete*

explanation [47]. Effective explanations provide appropriate and usable information, contingent on publics' existing knowledge and goals [105]. Explanations need not always be neat, structured, and comprehensive representations of complex entangled predictive systems [136].

Good enough explanations, therefore, are slow, ongoing or continuous and partial. Their goal is not to provide a one-time snapshot of how the predictive systems work in their entirety. Instead, good enough explanations may last several interactions and are intentionally partial.

Atlas of Surveillance [137], developed by the Electronic Frontier Foundation and the University of Nevada, Reno Reynolds School of Journalism, provides transparency into the technologies used by various police departments in the United States. Volunteers, slowly, but steadily, crowdsource information from media posts, press releases, government meeting agendas, and news articles to bring together regularly updated information. Today, as they report, they have over 10,000 data points that have been used to study how policing technologies are growing. Even though this information is incomplete and continuously growing, it has been highly impactful [138]. It has provided important context to several research studies [139] and has been good enough to launch investigations into the use of technologies by specific law enforcement agencies [140].

2.5.4 Actionable

Participants describe meaningful explanations to be a means to an end, rather than an end themselves. Explanations serve as starting points for severals purposes including but not limited to: accountability, discussion, reflection, assessment, and regulation. Researchers have critiqued the ideal of transparency alone [141]. They see the potential of transparency in supporting informed citizens to act [113] and impact the governance of technology companies through their roles as shareholders and consumers [142, 143]. The ability to act is key to effective explanations [98].

As such, good enough explanations are good enough if they are surrounded by mechanisms that allow social groups to act. They serve as precursors to other goals. The design of good enough explanations then includes the design of systems and infrastructures that promote action [144].

One promising grounds-up initiative for oversight over police technologies is called Community Control of Police Surveillance (CCOPS) [145]. This movement aims to provide ongoing transparency and control to community members over if and how city agencies acquire or use surveillance technologies. The adoption of CCOPS ordinances, and the transparency that it has provided, has resulted in the ban of oppressive technologies, such as facial recognition, in several states including California and Massachusetts [146]. In New York City, however, the adoption of CCOPS has yielded limited results in large part due to systemic barriers, which need to be overcome for transparency to be good enough. Knowledge of a lack of an explanation can also be good enough for specific contexts. Andrew Guthrie Ferguson, author of 'The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement' mentions in an interview that when the public asked questions about policing technologies, police officers did not have good answers [147]. Eventually, departments realized that predictive technologies may not be effective or worth the money. Explanations and transparency are only good enough in so far as they can inform the regulation of and control over the use of civic AI.

These qualities—situated, systemic, continuous and partial, and actionable—come together to formulate 'good enough explanations'. The qualities are not mutually exclusive or exhaustive. The questions asked by the participants and the related qualities I articulate, highlight that the process of creating an explanation for a civic predictive technology can be more important than the veracity of the explanation itself. Ultimately, this study defines good enough explanations as *continuous and partial processes* that allow *diverse publics* to engage with *features of predictive systems* and assess such systems in relation to their communities in *service of their goals*.

2.6 Conclusion

AI systems are not explainable in the binary of yes or no. Instead, this chapter attempts to conceptualize a 'good enough' explanation of civic AI. An understanding of such explanations is informed by the interviews I conducted with people who I believe to have a stake in the workings and effects of civic predictive tools such as academics, journalists, and leaders in civic organizations and neighborhood associations. I argue that good enough explanations arise through processes that allow diverse publics to assess features of predictive systems in terms of the effects on their communities.

In conclusion, this chapter contributes to existing XAI and AI transparency scholarship in the following ways: (1) it focuses specifically on studying public needs for effective explanations of civic AI systems, (2) it foregrounds the perspectives of non-tech experts and local publics who may have a stake in the design and use of civic AI systems, (3) it organizes empirical and theoretical knowledge to define what effective public explanations could look like.

My next step in this dissertation is to study systems that can support the creation of situated, systemic, continuous and partial, and actionable explanations. To that end, and motivated by research that presents the *form* explanations may take as an influential component of public explanations, I explore if and how participatory mapping can be used to study and create good enough explanations for diverse publics. In the next chapter, I demonstrate the affordances offered by participatory mapping as an explanation technique, followed by an account of the participatory mapping workshops I conducted in the course of this dissertation research.

CHAPTER 3

PARTICIPATORY MAPPING

3.1 What XAI Can Learn from the "Ghost Map"?

As I have already discussed, Human-Computer Interaction (HCI) and Machine Learning (ML) researchers have proposed several techniques for explaining and auditing algorithms and advancing algorithmic justice. However, more recently, XAI researchers have presented the importance of the *form* of algorithmic explanations and its effect on the understanding of AI systems [84]. Visualization experts and critical studies scholars have started employing visualizations as tools to explain and understand algorithms. I draw inspiration from this work to explore if and how *mapping* can be employed as a technique to study and design good enough explanations. Below, I summarize some existing works that draw on visual techniques to explain algorithms and their limits. Next, I explore the opportunities presented by *maps* by drawing on the work of Dr. John Snow, a physician-geographer, who traced the spread of cholera in the 1854 London epidemic through a map [148], popularly termed the 'Ghost Map' [149]. While these limits of XAI and affordances offered by maps to address them are not exhaustive, they are a good starting point to explore if and how maps can serve as useful tools to explain geo-spatial algorithms.

3.1.1 Visual explanations of AI

Integrating XAI and visual design can make customer-facing explanation interfaces more usable and readable [54]. The effects of static and interactive visualization techniques of white box (where a model's inner working is shown) and black box (where input and output variables' relationships are shown) explanations on user comprehension have been investigated [54]. It was found that white-box interactive explanations were most effective

in increasing user understanding but were worse than black-box explanations in increasing user confidence in their understanding. This may be a result of the complexity and cognitive overload of white-box explanations. An understanding of human cognition and decision-making capabilities is now being used to develop frameworks for explaining algorithms [49]. Researchers have also attempted to visualize ethical frameworks in order to make them more accessible to users who may not be familiar with ethical principles and terminologies [150]. However, more work is needed to design accessible white-box explanations of the design of algorithms.

Share Lab has used data visualizations to represent several aspects of algorithms including algorithmic labor, invisible infrastructures that surround algorithms, and the social and political relations that inform the workings of tech companies [151]. Kate Crawford's work titled 'Anatomy of AI' which displays several invisible aspects of labor, data, and environmental resources in relation to algorithms is another well-known example of visualizing algorithms [152]. These works draw great attention to the socio-political structures that surround the design of algorithms. However, they do not attempt to explain how the complexity of the world we live in is reduced to conform it to data practices.

Interactive visual analytics are being used to help data scientists better understand their systems through the design of tools such as Prospector [153], Gamut [154], Visual Auditor [155], and more. These tools visualize algorithms for controlled assessment and evaluation. More work is needed to incorporate real-life perspectives into algorithmic explanations.

Data Comics have been presented as a means to better report HCI and statistical analysis research studies and are being explored as visualization techniques to communicate research processes and practices in accessible and engaging manners [156]. Economist Julia Schneider and Artist Lena Kadriye Ziyal designed a comic series that explains what Artificial Intelligence is, its core properties, and the risks associated with its widespread deployment [157]. Explaining specific design decisions underlying the functioning of algorithms and their impact on the lives of its users still remains underexplored.

These visual approaches to explaining AI are highly innovative and serve as great starting points for furthering research at the intersection of information design and XAI. However, the methodologies underlying these approaches to explain algorithms remain limited across four primary dimensions as I describe in my previous work [158]. I propose that these limitations may be well addressed by one popular visualization method for cities—*maps*. The limits of existing explainable AI (XAI) approaches and the opportunities presented by maps to address those are listed below:

1. Accessibility: XAI methods can be incomprehensible for everyday users. On the other hand, maps are well understood—even treated with affection—by a broad spectrum of audiences.

2. Cultural reflexivity: XAI methods are commonly not representative of the social and political factors that shape algorithms and their designers whereas maps signify their own context of production through their visual languages that are culturally rooted.

3. Situatedness: XAI methods tend to be distant from situated real-world contexts and experiences of city inhabitants. Maps, however, draw on local knowledges of people and places in their making.

4. Design visibility: XAI methods focus on explaining the relation between input and output variables while disregarding how their representations of cities guide these relations. Maps reveal how the maker structured and segregated the city through its visual components.

This approach attempts to further the work of the scholars above and calls for the use of *mapping* to produce grounded explanations of the inner life of cities in ways that are accessible, culturally reflexive, situated, and provide visibility into their design. I discuss how *maps* can be effective in furthering such explanations and addressing the gaps in existing XAI research by drawing on the work of Dr. John Snow. I describe his work briefly in the section below.

3.1.2 The Ghost by John Snow: a brief description

Dr. John Snow, a pioneer in social mapping, had been studying the periodic cholera epidemics in London since the 1830s. He hypothesized that cholera is water-borne and is caused by the ingestion of contaminated water in contrast to the then-popular belief that cholera is caused due to poisonous air. In the 1854 cholera breakout, Snow set out to identify and prove the cause of the disease. To do so, he mapped the fatalities and their proximity to various water pumps in the form of bars and dots. He noticed that the majority of deaths were taking place around the Broad Street water pump. However, there were some anomalies. A major brewery and a workhouse were largely unaffected by the disease. Upon talking to the people who lived and worked there, Snow found out that they both had their own individual wells for water consumption and the brewery workers majorly drank beer instead of water. He also observed that there were cases of outbreaks in areas distant from the Broad Street pump. Later, Snow found out that the people living there were still consuming the broad street pump water either because it was their place of work or school or because they enjoyed the taste of that water more. Snow presented his map to the authorities and citizens upon which the city agreed to remove the handle of the Broad Street pump to prevent people from consuming its water. Eventually, the epidemic ended. London has not seen a cholera breakout since [148]. What can I, along with fellow XAI researchers learn from this historical example, as we seek to explain the algorithms that govern "smart cities?"

3.1.3 Algorithmic Explainability: Limits and Opportunities of Mapping

In the section below, I reflect on four strategies that maps have used to explain the complex inner lives of cities— accessibility, cultural reflexivity, situatedness, and design visibility. As a salient and well-known example of the explanatory power of maps, I will make use of John Snow's "Ghost Map" of London described above.

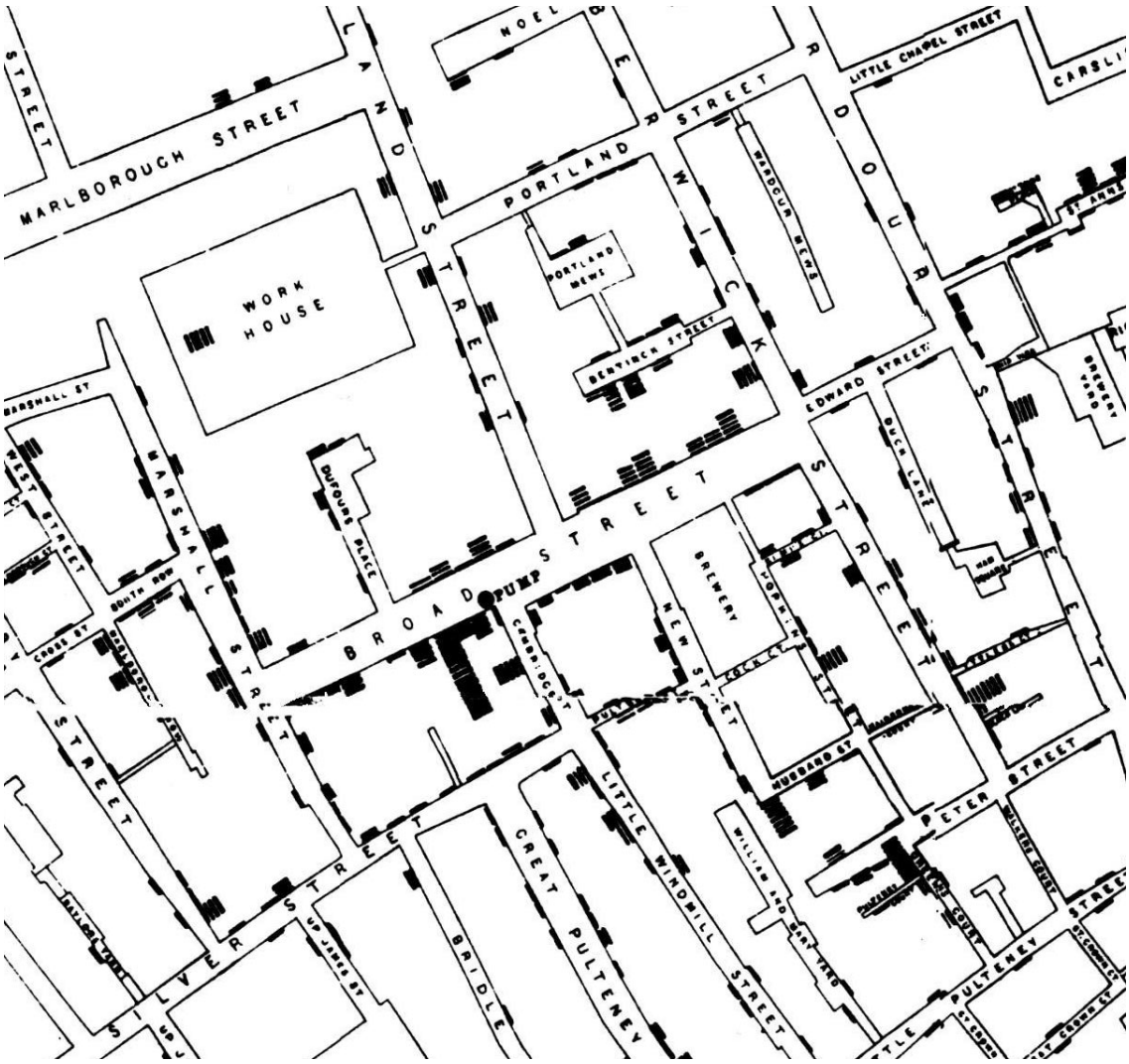

Figure 3.1: John Snow's map that traces the spread of cholera deaths during London's 1854 cholera epidemic

Accessibility

Before John Snow's work on the Ghost Map revealed that Cholera is spread through drinking water, the London water distribution board was convinced that it was by something in the air, a miasma. They refused to believe the research presented by John Snow until he was able to supplement his research with a visual, easily understandable, and detailed representation of the precise spread of the epidemic. Snow gave spatial form to the General Registrar's data about cholera deaths. He converted one-dimensional data organized by date

of death to two-dimensional data organized by proximity to water sources. This conveyed the cause-effect relationship between cholera infections and proximity to the Broad Street pump efficiently [159] and clearly. The accessible nature of the map allowed many others to not only understand the map but to reproduce it [149]. It neatly captured the intricacies of the complex urban phenomenon that spread the disease and convinced the general public and authorities of its claim [160, 161]. According to Johnson Stevenson, the map helped make sense of a phenomenon (microorganisms invisible to the naked eye) that previously "defied human understanding" [149]. Ultimately, the map served as an excellent example of a user-friendly visual explanation. Popular XAI approaches are inaccessible to everyday users. Maps present an opportunity to explain algorithms in a comprehensible manner.

Reflexivity

Visualizations bear the burden of being interpreted as a single objective reality of the world. However, this can be overcome with thoughtful consideration. John Snow, in his mapping and supplementary description, does not remove himself from his map. Rather, he carefully documents the possible errors that may have occurred in the data collection and presentation processes alongside the reasons for the possible errors. First, he clearly shows the contrast between deaths in houses near Broad Street pump and the brewery and workhouse nearby that remained unaffected. This leaves room to question the cause-effect relationship that Snow hypothesized. Second, in accompanying texts he lays out how some data he received from the registrar's office was missing house numbers and so could not be visualized [160]. While this map itself is a reflexive artifact in the sense that it captures Snow's acknowledgment of his partial perspective through his mapping and descriptions, it also is an artifact that promotes reflexivity in the reader by challenging their stable perspectives. In its form, it communicates the time period it was created in and for what purposes, the region and scope of analysis, and what London looked like at the time.

Situatedness

The data represented in the Ghost Map was informed by the everyday lives of the residents of Broad Street and the rest of London. To be able to design the map, Snow needed a fine-grained knowledge of the people of London. He conducted extensive interviews to understand people's movement patterns, sanitary conditions, water consumption practices, etc. He also collaborated with local people such as those who attended the sick. Even though the Ghost Map presents a bird's eye view so that the reader can observe patterns, it builds on situated knowledges of local experiences [149]. For example, Snow investigated why a brewery close to Broad Street was unaffected by the disease. He found out that it was primarily because the brewery workers mostly drank beer instead of water which saved them from the disease [160]. Alongside considering the spatial distribution of the water sources, Snow also took into account the time it would take when walking along the turns of the city for a house member to reach various water sources. This further exemplifies the incorporation of situated elements, including temporality, that inform the design and analysis of the Ghost Map.

Design Visibility

The Ghost Map followed and represented a clear hypothesis: Cholera spreads with the consumption of contaminated water. To analyze this hypothesis, Snow clearly displayed only two major components on the map— the water pumps in London and cholera-related deaths. He had a clear causal relationship in mind that he hoped to investigate and that was made clear to the reader [160]. The clarity of the map also provided space for critical questions such as how the selection of time periods (aggregating cholera deaths over a week or a day) and boundaries drawn (deaths in a house or a block or another abstract segregation) affects our understanding of the spread of disease [160]. Tufte critiqued John Snow's dot maps by arguing that visualizing deaths as dots gives little information about the population density of different areas. That is, deaths would likely be more in a densely

populated area whether or not a water pump was the source of the disease. Had it not been for the map of Snow's analysis, Tufte would not have been able to critique and investigate the basis of the claim made by Snow. Mapping how algorithms represent cities provides similar opportunities for critique. For example, when calculating the safety of a location, does the algorithm aggregate the crime rate without normalizing it by the population density of an area? If yes, what are the impacts of such design decisions?

3.1.4 What can explainable AI learn from the Ghost Map?

In the sections above, I demonstrate the usefulness of *mapping* as a technique to provide grounded explanations of spatial phenomenon in cities. I identified four strategies that conventional maps use to uncover the complex inner lives of cities. These strategies, as I argue in my work [89], can prove helpful when explaining geo-spatial algorithmic systems: *accessibility, cultural reflexivity, situatedness, and design visibility*.

Comprehending our urban environment by "making public data public" has been considered of utmost importance by many architects and city planners [162]. While I have focused on the case of the influential Ghost Map by John Snow, many other cartographers, visual and information designers, and urban planners present similar noteworthy works that demonstrate the useful qualities of maps in explaining cities. Here are a few other references that XAI researchers might yet explore:

Accessibility: Popular XAI approaches are inaccessible to everyday users. Maps present an opportunity to explain algorithms in a comprehensible manner. In 'The Image of the City' [163], the proto-"city designer" Kevin Lynch describes how he and his students at MIT identified five constitutive elements that city dwellers use to understand the places they live in: paths, edges, districts, nodes, and landmarks. The implication is that these five elements can form the core of "legible" city design and mapping. With the longest print run of any book by MIT Press (more than eighty years), Lynch's work still serves as a useful tool for thinking about how the complex form of cities can be accessible.

Reflexivity: XAI methods do not explain the social-political contexts that shape algorithmic design. This limitation can be addressed by the calls for critical reflexivity in the mapping and representation of cities in fields such as 'Critical Cartography' [164]. Such representations can challenge the status quo in a number of ways. Challenging western positivist cartographic techniques, some mapmakers are creating alternative forms of mapping that foreground indigenous forms of spatial knowledge [165]. Another similarly subversive map designed by Laura Kurgan and her collaborators called "Million Dollar Blocks," draws critical attention to unequal practices of incarceration that specifically target Black neighborhoods in New York City (USA). The map draws attention to a few low-income city blocks where millions of dollars are being invested—not for public health or education—but to remove the inhabitants and put them behind bars in the name of creating a safer city [166].

Situatedness: XAI methods tend to be distant from situated real-world contexts and experiences of city inhabitants. On the other hand, several artists and scholars have experimented with participatory mapping projects, that allow participants to make meaningful connections between authoritative data, such as census or censor readings, and their own local knowledge of the places they live in. One such example is the Atlanta Map Room Project [167], an expansion of the St.Louis Map Room [168], that brings together local community members to explore and represent the relationships between civic data and lived experiences. This grassroots mapping effort problematizes the "objective" and "stable" nature of big data and grounds data and its limits in contextual experiences of space.

Design Visibility: XAI approaches limit their scope to explaining the relation between input and output variables while disregarding how algorithms construct cities in their design. Decisions about what one chooses to represent on a map have a significant impact on who is affected by city design positively or negatively [169]. Who is on the map or off the map affects who is included and who is not [170]. While maps construct reality through their representations, they also, through their very form, explain to us their constructed

reality, leaving room for critiquing and improving said reality.

Our goal here is take inspiration from the cultural history of visual design to design AI explanations. In this dissertation, I employ the above-mentioned affordances of maps to explain public safety algorithms to city inhabitants. I employ mapping as a technique to explore and explain how algorithms and the spatial aggregations they may create are grounded in the historical, economic, political, and social contexts of cities.

Gupta et. al have shown how changes in spatial partitioning of cities for aggregation of data can have major effects on algorithmic results. For example, they demonstrate how the Gini index (a measure of spatial inequality) values change as one modifies the scale of calculation [171]. Given the impact of spatial partitioning on algorithmic outcomes and the need for algorithmic transparency in such cases, I conduct participatory mapping workshops with diverse groups to visualize algorithmic partitioning on maps, superimposing these distributions with other data layers that represent historical and contemporary segregations such as red-lining maps to ground the partitioning in historical politics. My hope is that this could help us evaluate if and how algorithms may reinforce or challenge existing city segregations. I describe these workshops in the section below.

3.1.5 Limits

Even though I plan to use mapping to study and question spatial injustices, historically, mapping has also been used to propagate discriminatory agendas [172]. Maps may present distant and stable representations of reality which may favor the perceptions of one social group over another. However, with advancements in critical cartography and GIS, there are now new ways to question the ideologies and assumptions embedded in "objective" maps. Scholars are also exploring participatory mapping as a technique to problematize the positivist representations of cities and open up mapping for pluralistic exploration and critique [173]. I present mapping as an exploratory tool for advancing grounded XAI research while acknowledging its limitations.

3.2 Workshops

3.2.1 Workshop Setup

Figure 3.2: Participatory Mapping setup: Projection of a map on a table on which participants can draw

The section above details our motivations to employ participatory mapping as a tool to explain and question geospatial civic AI. To conduct participatory mapping sessions, I used an existing tool named MapSpot [174], a part of the Atlanta Map Room Project [167]. This project is an expansion of the St.Louis Map Room [168], that brings together local community members to explore and represent the relationships between civic data and lived experiences. This grassroots mapping effort aims to problematize the "objective" and "stable" nature of big data and grounds data and its limits in contextual experiences of space.

In our mapping sessions, participants gathered around a mapping set-up that included a table covered with drawing paper and a short-throw projector turned 90 degrees to project

a digital map on the table. The projection acted as a guide for participants to map spatial components related to AI systems. On this paper, participants mapped their place-based experiences of safety and how they may relate to spatial algorithmic predictions. The projected digital, as part of the Mapspot tool [174], also allowed us to overlay data layers onto the mappings of participants including demographic data such as census data or historical data such as red-lining maps (see fig. 3.3). Such superimposition of grounded local data and institutional data allowed the group to understand if and how space, and the people who live there, would experience predictive technologies differently.

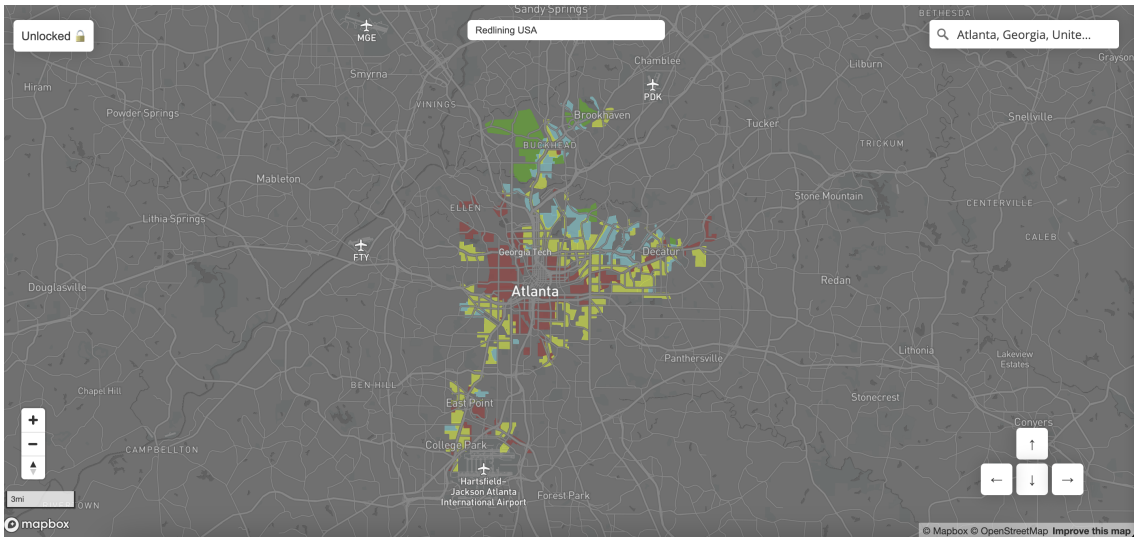

Figure 3.3: Mapspot website displaying the red-lining maps data layer

3.2.2 Pilots

I conducted two pilot mapping sessions in April 2023 that focused on the design and impacts of place-based predictive policing tools. These sessions were conducted as part of the Digital Media Demo Day [175] and the GVU Research Showcase [176] on the main campus of the Georgia Institute of Technology in Atlanta, GA. These demos were done in an exhibit setting and the mapping sessions typically lasted between five and twenty minutes with participant groups ranging between one and five participants. In total, approximately twenty people actively participated.

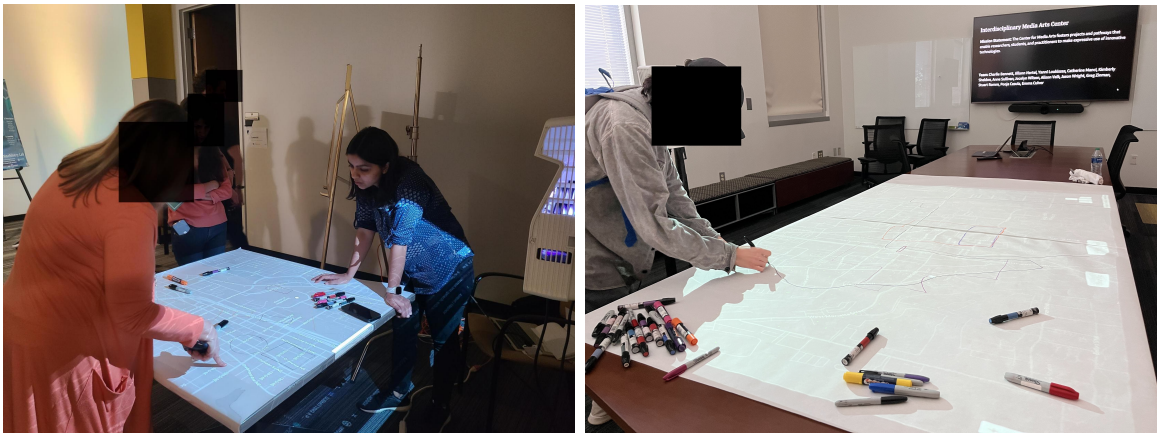

Figure 3.4: Pilot study sessions conducted as part of Digital Media Demo Day (left) and GVU Research Showcase (right) in April 2023

Participants included Georgia Tech alumni and other industry, government, academic, and civil service individuals who were interested in learning more about the projects and research happening at Georgia Tech. Demographic details of the participants were not collected due to the format of the interactions. However, by virtue of the places and networks where these events are advertised as well as the voluntary sampling of people, it can be estimated that the audience was more technically savvy or interested in learning more about how humans and computers interact.

The goal of the pilot studies was to understand if a participatory mapping activity shows potential to investigate the research question this dissertation asks: *How can the Explainable AI (XAI) community support public understanding of the spatial workings and effects of civic predictive systems?* . To do so, participants engaged in a speculative exercise where they imagine and discuss how a predictive system may categorize the places they know and live on on the map and why.

Through this exercise, I found that participants were able to ask critical questions about AI systems and their spatial effects. Some examples include: inquiring about the impacts of over-reporting and under-reporting of crime and on AI predictions; questioning the ends towards which these predictions are used (increasing police presence or systemic change);

or discussing predictions for white collar crime data vs blue collar crime data.

Participants also reflected on participatory mapping and speculative personation as a methodology in supporting such critical thinking. A participant said:

> I think the interactive piece helps ground this in their own experience to think about ok.. where do I live.. where my actual life took place in this area. So I can actually think about streetlights and bus stops and things like that.

It was considered an "*excellent way of getting the conversation and pulling it out of people*". Another aspect participants discussed was the embodied nature of the personation experience and expressed discomfort in making algorithmic predictions. For instance, a participant said:

> I think for me forcing me to draw areas that I think would be flagged was like really fascinating.. I did not want to do that cuz I was like.. there was some uncomfortableness in doing that.

Speculative personation using maps allowed participants to reflect on the moral responsibility of making high impact predictions, a responsibility that can never be felt by AI systems as they cannot feel [177].

These early pilot studies showed promise in the use of participatory mapping and speculative personation for collectively understanding AI systems and the broad socio-technical assemblages that surround them. Following these demo pilots, I conducted a workshop with Georgia Tech students as a dry run leading up to five primary workshops that inform the rest of my dissertation. I report on these workshops below.

3.2.3 Workshops

Five workshops were conducted in person in Atlanta, Georgia located in the Southeast United States. The groups who participated in the workshops were invited through one of

the following ways: (1) they were recruited through open calls for participation which were shared with numerous non-profits and civic organizations via email, (2) they had existing relationships with the authors and authors reached out to ask if they would be interested in participating, and (3) they were familiar with the ongoing work of the authors and reached out to participate. The workshops were conducted as one-time 90-minute workshops in either the offices of participating organizations or the university campus of the authors. I led the moderation of every workshop with organizational and documentation support from colleagues. The workshops were audio and video recorded. The study was approved by Georgia Tech's IRB.

Before each workshop, a short survey was shared with the participants. The survey served two primary purposes: (1) it confirmed participation through an informal registration process, and (2) it provided workshop moderators with background information that was helpful for unique workshops (see Appendix B). The survey asked participants about the Atlanta neighborhoods they were familiar with so I can focus our conversation on those areas, their goals for participation, their familiarity with AI tools, and any existing questions, opinions or concerns they had about AI systems. This information helped me, and my fellow moderators, in understanding the explanation contexts we would be moderating.

The workshops started with participants engaging in what I call 'speculative personation'. They were asked to make spatial predictions as an AI system would. They identified places on the map that they were familiar with and believed a predictive tool may identify as crime hotspots. They were then asked to reflect on their predictions by considering their reasoning and effects. The workshops followed a loose protocol that guided such reflection. Participants contemplated an AI system's prediction goal, data type, data source and amount, spatial aggregation of data, and prediction effects in relation to their own speculations. In doing so, moderators prompted participants to consider the socio-political qualities of places they identified in relation to these technical aspects of the AI system. A focus on specific locations revealed spatial, social, political, and historical characteristics

relevant to the technical workings of the predictive tool. The protocol was documented as a toolkit and shared as a take-away artifact with the workshop participants to support their questioning of AI systems beyond the workshops (see Appendix C). In Workshop 3, I introduced another page to the artifact that provided information on other AI powered public safety tools in the US, if and when were they used in the city of Atlanta, and a one-sentence description of media articles that document their real-life impacts. The need for this information was felt in the earlier workshops as participants discussed public safety AI more broadly. It is essential to note that this structure of the toolkit was not rigid and merely acted as a guide to the conversation while allowing the group to deviate into explaining and questioning components of AI systems that were most relevant.

I, with the support of my colleagues, conducted workshops with a Police Reform Group (Workshop 1 or W1), City Planners (W2), Neighborhood and Civic Representatives (W3), a Community Development Organization that funds technology innovation projects (W4), and Educators part of an education non-profit (W5). In chronological order, I describe these workshops below.

Table 3.1: Participatory Mapping Workshops conducted as part of this dissertation

Workshop	Group	Participant Count	Timeline
Pilots	Georgia Tech Demo Days	≈ 20	Apr 2023
W1	Police Reform Group	12	Oct 2023
W2	City Planners	7	Nov 2023
W3	Neighborhood and Civic Representatives	7	Nov 2023
W4	Community Development Organization	11	Jan 2024
W5	Educators	8	Jan 2024

Workshop 1: Police Reform Group

W1 invited twelve members of a police reform group that offers supportive services to those experiencing extreme poverty, problematic substance use, or mental health concerns, with the goal of reducing their arrest and incarceration rates. This workshop was conducted

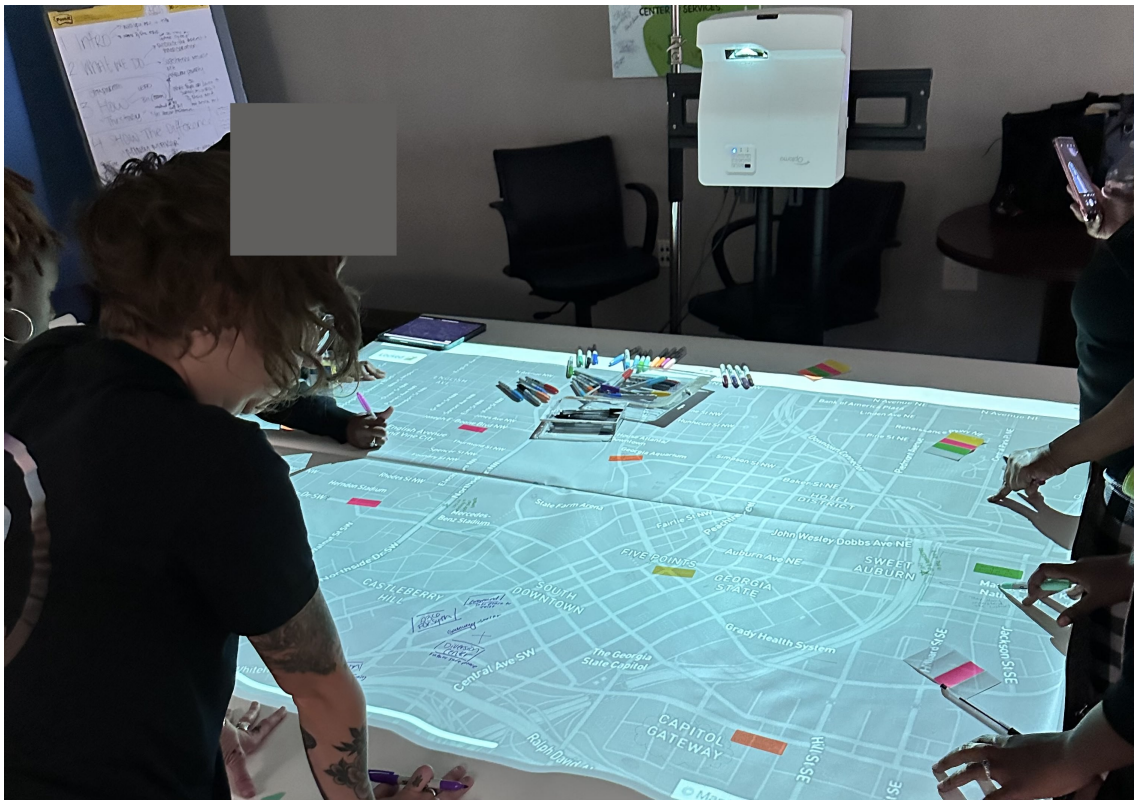

Figure 3.5: W1 Participatory Mapping with police reform organization

in October 2023 in the offices of the partnering organization in Atlanta. Before the workshop, I conducted a 30-minute call with the director of the group to understand their goals, motivations, and expectations of the workshop.

In their survey, participants reported that they want to learn about the new and advanced methods in policing, what is AI and how AI will be used to promote public safety, how to use these tools responsibly, and how it can help them better serve their communities. A few of them explicitly stated that they had concerns about the use of AI in policing, especially because of the effect of racism on policing, and saw this workshop as a place to discuss those concerns. Participants described themselves as having 'none' to 'very little' knowledge of AI. Two mentioned their use of ChatGPT and another talked about their knowledge of policing technologies like surveillance cameras.

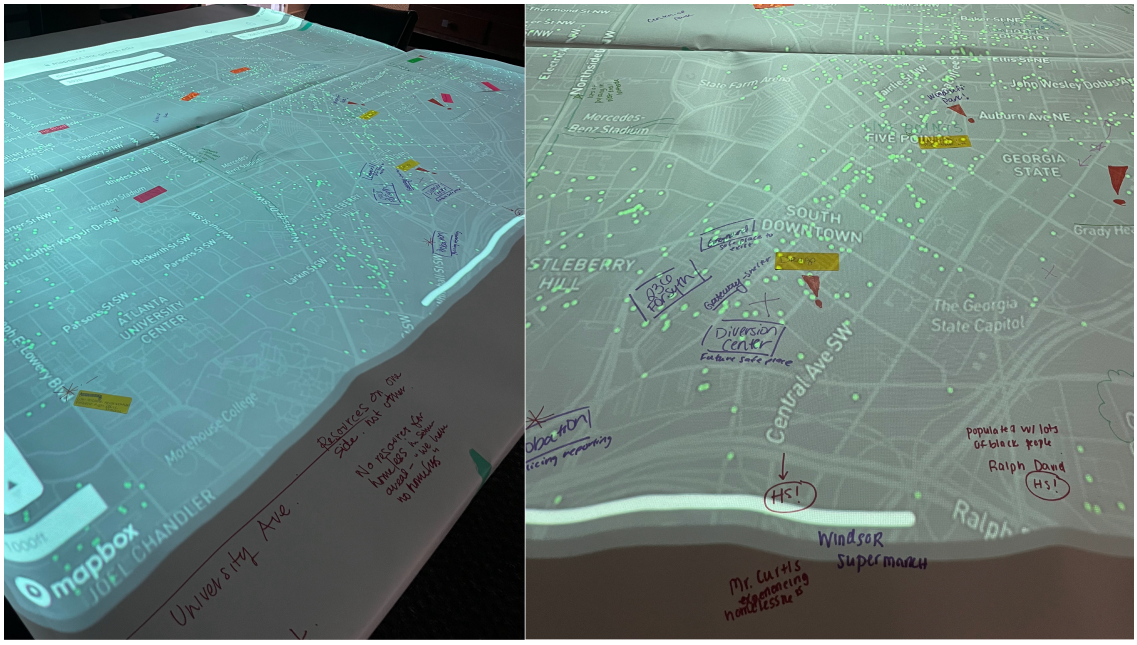

Figure 3.6: Participatory Maps Close-ups

Workshop 2: Urban planners

W2 invited seven members of a regional civic planning agency in Atlanta. It supports the leaders of Atlanta in using data-driven methods to plan, invest, and addresses critical issues for the betterment of city's collective future. The workshop took place in the first week of November 2023 in the offices of the organization based in Atlanta.

Participants were interested in learning about how AI can be used for the betterment of the city and promote equity. They mentioned their roles as urban planners and designers and sought knowledge of civic AI tools to skillfully engage with these tools in their professional work. Five participants described themselves as being familiar with AI through podcasts, readings, use of ChatGPT, their own work, or the work of those around them. The group seemed excited about the opportunities offered by AI but wanted to learn about how to design, deploy, and use sch tools in an equitable manner.

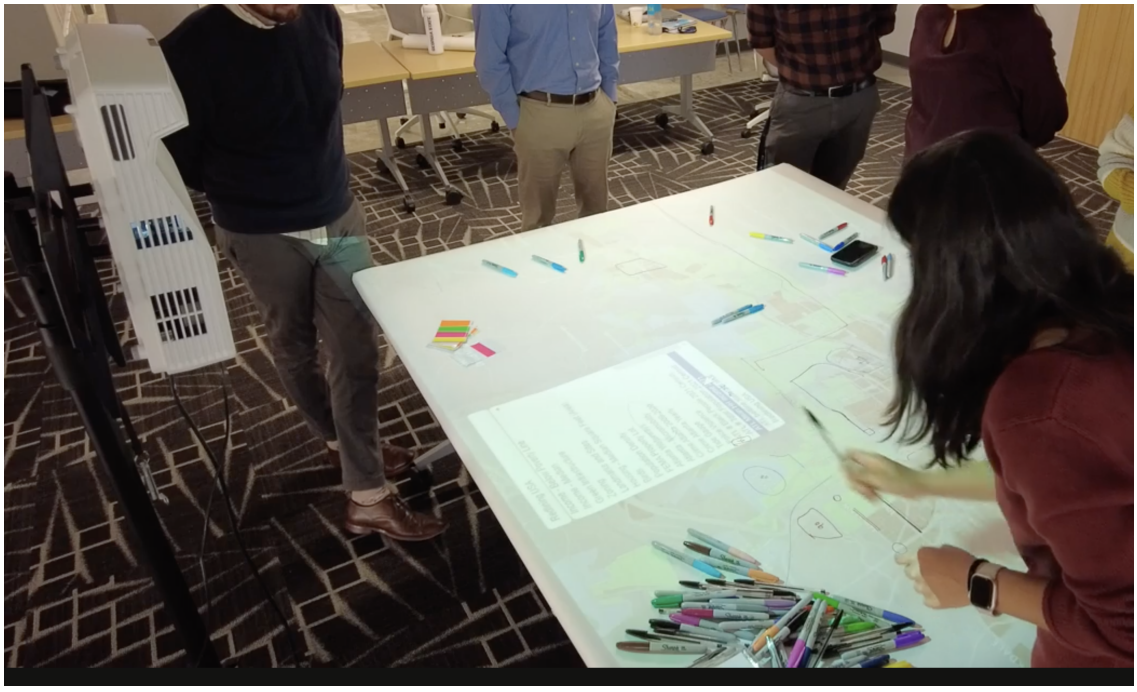

Figure 3.7: W2 Participatory Mapping with city planners. Selecting data layers.

Workshop 3: Open Call

Unlike the rest of the workshops, W3 was an open call workshop publicized via a webpage, email, and flyers with neighborhood association leaders, board members, and other neighborhood and civic groups. Folks were encouraged to share and publicize the workshop details to anyone in their network who they deem would be interested. The workshop received seventeen registrations of which seven people participated in the workshop. The workshop was organized on the Georgia Tech campus in mid November.

This workshop saw a wider range of people. The participants included researchers and practitioners from both academia and civic organizations for neighborhood improvement. One participant worked in violence prevention in Atlanta and another was leading a neighborhood public-private partnership aimed at making spaces better for citizens. In the survey, some described themselves as "Pretty familiar with usage" of AI tools and others as "Not very familiar" with AI. Some had friends who worked in AI and others had used tools with predictive features. Participants were interested in how AI could be used for

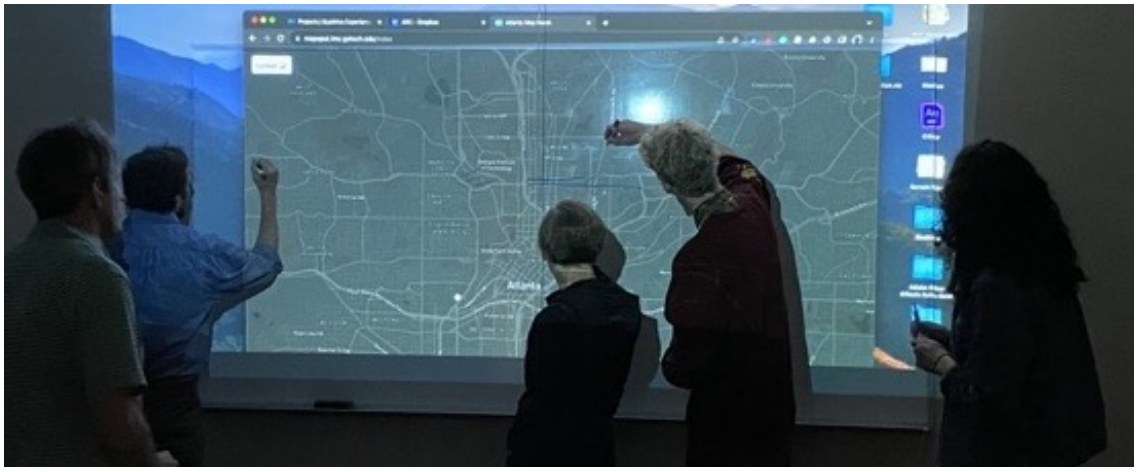

Figure 3.8: W3 Participatory Mapping with civic and neighborhood organizaton representatives on the wall

civic purposes in responsible ways. Two participants were aware of the biases that can creep into AI systems and considered the workshop a space to learn and talk more about it. This workshop was met with a technical error where our short-throw projection on the table stopped working and I, along with my co-moderator, improvised to a wall displayed map.

Workshop 4: Community development organization

W4 brought in eleven members of an organization that leads, manages, and funds local, community-centered projects across Georgia. This workshop was organized as part of an annual retreat for the organization members where members gathered to discuss strategy for the rest of the year. The organization director met with me before the workshop and considered the workshop timely and useful for members to understand how AI systems may affect the city. Such an understanding, the director hoped, would support their work in civic development for the smart city. The workshop was organized in their offices in early January 2024.

Most people were participating as the workshop was included as a part of their team retreat. They were interested in incorporating the learnings from the workshop into their

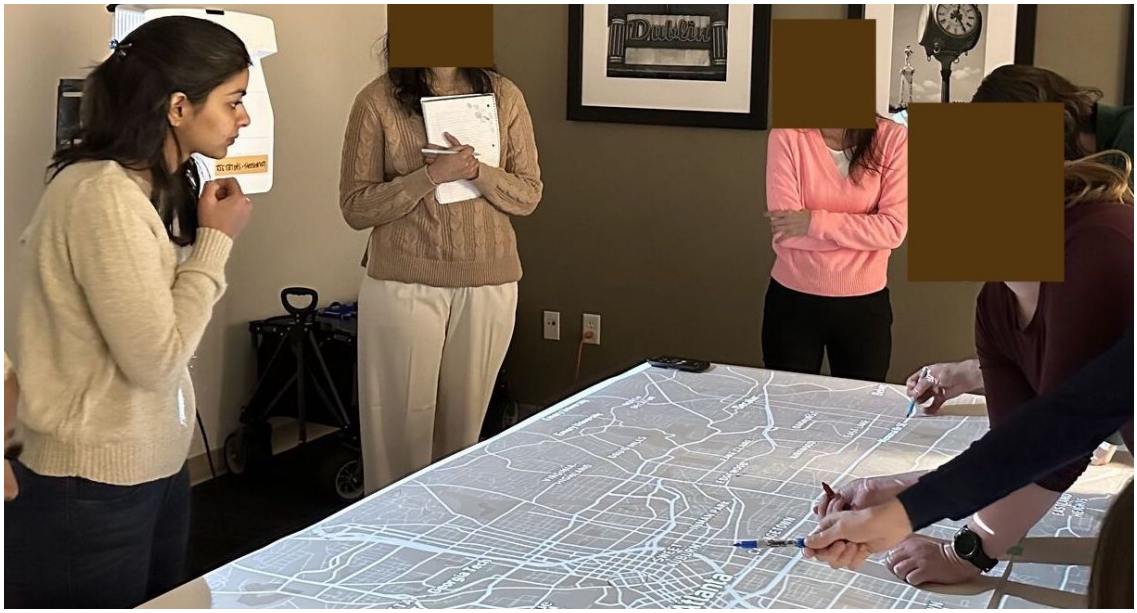

Figure 3.9: W4 Participatory Mapping with community development organization

work. Participants mentioned different predictive tools they were familiar with. Three participants mentioned variations of ChatGPT, Claude or Bard while some talked about Google maps and Google translate. Overall, the group presented themselves as being familiar with AI through various articles they read, tools they have used, collaborations with AI researchers, or their own work in smart cities. One participant described their interest in learning how to make these tools transparent to the public. Another talked about the privacy and safety of these tools rural communities.

Workshop 5: Non-profit of educators

W5 invited a group of eight teachers who came together as a non-profit aimed at forming meaningful relationships and collaborations between educators. The non-profit leaders offer experiences to support teachers in their work and, with this workshop, hoped to come together to learn more about AI and its effects on cities. The workshop was organized on the Georgia Tech campus in late January 2024.

Participants wanted to learn how to responsibly incorporate AI in their teachings, especially with the rise in generative AI and its effect on education. Two teachers taught

Figure 3.10: W5 Participatory Mapping with educators

geography, another taught social science. One participant said they had "not given AI a great deal of thought", while another participated in a school committee to discuss AI use in high school education.

Three of the five workshop partners posted about this work on their social media channels reflecting on their experience and demonstrating the impact the workshop had such as:

> "As we look to the future of these technologies, we have the potential to enhance community well-being and resource distribution, but only if we fully understand their potential and limitations." (W4)

> "Our engagement [through the workshop] not only broadened our understanding of the power of maps and the ethical implications of AI but also served as a reminder of our role as educators in shaping future perspectives." (W5)

These workshops together inform the rest of the work discussed in this dissertation. I analyse these workshops through two primary methods (1) situational analysis and (2) thematic analysis and memo-ing which I discuss in Chapter 4 and Chapter 5 respectively.

CHAPTER 4

SITUATED EXPLANATIONS AND RELATED CONTEXTS

We established in Chapter 2 that good enough explanations are situated. This chapter explores this quality in greater depth and asks: *How can we understand the explanation needs of diverse publics as situated in their broad-ranging contexts?* As the emerging field of Human-Centered Explainable AI (XAI) aims to design algorithmic explanations for diverse users, they consider user needs to vary along two dimensions: (1) users' AI literacy levels and (2) users' roles as stakeholders in the AI ecosystem. While considering the relationship between users and predictive tools, defined by their literacy and roles, is necessary for the design meaningful explanations, it is not sufficient. This chapter demonstrates how publics' explanation needs emerge out of broader relations with algorithmic contexts. To do so, I report on the five participatory mapping workshops detailed in the previous chapter. These workshops with diverse groups were aimed at collectively questioning and understanding AI systems. I analyze how publics relate to the contexts around them as they seek to understand predictive systems. I find that publics' explanation needs emerge not just out of their relation with the predictive tools but also through their relations with predictive domains, predictive subjects, and predictive backdrops. As such, I urge human-centered explainable AI designers to consider these relations as they design systems to support public understanding of AI.

4.1 Introduction

For the public to be able to effectively use, manage, oversee, assess, or challenge pervasive AI systems, they must understand the workings and roles of AI in their lives. Earlier efforts to explain AI focused on AI developers to support them in debugging their models. However, more recently, policymakers and researchers are calling for human-centered Ex-

plainable AI, i.e. creation of explanations in ways that meet the needs of users, including everyday publics.

Current human-centered XAI work considers explanation needs to vary based on users' relations with predictive tools across two dimensions: (1) AI literacy: the pre-existing knowledge people have about AI tools, and (2) Roles: their position as a stakeholder such as user, auditor, policy maker, etc. Ongoing investigation into both these dimensions is necessary to understand how users' explanation needs come to be. However, it may not be sufficient. As Nicenboim et al. argue, explanations are contextual [178]. Such contexts can be broad-ranging and include rich and nuanced relations with other aspects of the algorithmic ecosystem. These contexts, as Paul Dourish argues, cannot be pre-supposed. They are not defined by a set of stable conditions that exist outside the action of people. Instead, they emerge out of the course of action. As such, they can only be reflected upon in hindsight. I term the contexts that emerge in the process of explaining AI— 'explanation contexts'. Actors, materials, knowledges, and relations that become relevant to questions and understanding AI become part of continuously re-negotiated explanations contexts. Such a holistic study of explanation contexts remains understudied. To that end, this chapter asks: *How can we understand the explanation needs of publics' as situated in diverse explanation contexts?*

To answer this question, I draw on the five participatory mapping workshops I detailed in Chapter 3. These workshops invite participants to engage in the processes of questions and knowing place-based predictive policing systems. I reflect on these workshops to understand how explanation contexts reveal themselves and how users' explanation needs emerge from dimensions beyond what has been identified by existing literature (i.e. literacy and roles). I employ Clarke's 'situational analysis' [179] to identify how publics relate to algorithmic contexts as they seek explanations.

We find that publics relate to algorithmic contexts through (1) the predictive domain: the service domain that an AI model becomes a part of— in our case policing, (2) the

prediction subject: the people or places that are subject to predictions—in our case neighborhoods, (3) the predictive backdrop: the local and global environment that surrounds predictive systems, and (4) the predictive tool: the tools and models that make predictions. I argue that publics explanation needs emerge out of their personal or communal relations with these elements. While existing research in user-centered XAI has majorly focused on the fourth element—relation to the predictive tool through user literacy or role— I suggest that understanding some of these other relations can help XAI researchers provide meaningful explanations situated in the lives of diverse publics.

In what follows, I provide a summary of human-centered XAI research and how they understand algorithmic context in section 2. Next, I describe my use of situational analysis and situational maps to analyze the workshops in section 3. In section 4, I report on the explanation contexts that emerge in three of the five workshops I conducted. Lastly, I propose a framework that can help XAI designers in considering the broader relations through which users' explanation needs emerge.

4.2 Background

4.2.1 Explanations needs for diverse AI expertise

Existing knowledge and understanding of AI have been used as a metric to conceptualize user-centered explanations. Researchers relate low literacy levels in non-expert users to higher unwarranted trust in AI systems, thereby calling for advancements in user understanding of AI [180, 181].

Users' AI expertise is understood in many different ways. Some works clearly define user expertise using scales or metrics. In their survey, Hohman et al. categorize the audience of XAI as 'Model Developers and Builders', 'Model Users', and 'Non-Experts' with decreasing levels of AI expertise [182]. Similarly, Yu and Shi report on the goals of users based on whether they are beginners, practitioners, developers, or experts [183]. Mohseni et al. focus specifically on data literacy and add 'Data Experts' as a category to define user

expertise [184].

Some researchers have used technical means to codify user expertise into their explanation platforms. Rong et al. propose a framework called Image Classification Explanations or I-CEE encoded in an explanation tool, which provides explains AI to users by providing them with a sub-set of training data based on their expertise. The user expertise is categorized as an m-dimensional vector that informs the explanation provided to them [185]. Other researchers have however chosen participatory ways to investigate what to explain and how to explain such that users' mental models are more closely aligned with expert mental models [106].

Ehsan et al. emphasize the need to consider 'who' is attempting to understand AI, arguing that users' underlying characteristics, such as their AI background, motivate their explanation needs. They demonstrate how groups with different understandings of AI, such as Computer Science students versus Amazon Mechanical Turk (AMT) participants, use diverse heuristics and appropriation techniques when engaging with explanations [180].

4.2.2 Explanation needs for diverse user roles

Depending on the type of stakeholder a user is, and what their functional role is in relation to the predictive tool, they may have different explanation needs. Tomsett et al. propose six different roles namely: Creators, Operators, Executors, Decision-subjects, Data-subjects, and Examiners. They speculate explanation needs for these personas such as optimizing models for creators or contesting decisions for decision subjects [186]. Preece et al. expand the binary distinction of developers and users to also consider the explanation needs and motivations of theorists and ethicists. They suggest that a 'layered' approach to explanations can meet the needs of diverse stakeholders [187]. Hong et al. focus specifically on ML practitioners and identify explanation needs for the roles of model builders, model breakers, and model consumers. They emphasize that explanation needs vary with contexts and depend not only on stakeholder roles in relation to the predictive tool but also on

relations between workers in an organization [188].

These works demonstrated how explanations needs are conceptualized in relation to users' AI expertise and functional roles. They presuppose the explanation contexts of users in relation to their expertise and roles and design systems and tools for such contexts. Even as these approaches show promise for the design of user-centered AI explanations, how they conceptualize explanation contexts is reductive.

4.2.3 Towards a more holistic understanding of users' explanation contexts

The categorizations of user expertise and roles as described above remain limited. The knowledge of the audience is merely evaluated based on their technical expertise, disregarding other knowledges that may affect their understanding and explanation needs such lived experiences. Additionally, their explanation needs are singularly considered a factor of their presumably static roles [129]. In response, Suresh et al. present a framework to granularly describe stakeholders' knowledge and interpretability needs. The first part of their framework focuses on stakeholders' expertise and describes two dimensions: stakeholder knowledge (formal, instrumental, and personal knowledge) and the contexts in which this knowledge manifests (i.e., machine learning, the data domain, and the milieu). The second part describes stakeholder needs through a three-level typology: long-term goals (understanding the model, and building trust in it), shorter-term objectives, (debugging a model, or contesting a decision) and immediate and specific tasks (assess prediction reliability and detect mistakes). Their framework emerged out of a broad survey of interpretability and pedagogy literature [129].

Researchers have also highlighted the need to study user explanations not in an individual context but in relation to the contexts and communities that surround them. Kou and Gui draw on Activity Theory and present a framework of six components and interrelationships namely: subjects (peoples and platforms involved in explaining), tools (that mediate the process of explaining), division of labor (who is explaining), rules and com-

munity (circumstances within a platform), and objects (contextualizing AI) to seek socially oriented, systemic-oriented, and mechanism-oriented explanations [189].

While works on nuancing explanation contexts is very nascent, these works present some ways forward to develop a more holistic outlook of explanation contexts. I build on these works to define explanation contexts more precisely in this chapter.

4.3 Data Analysis

In this chapter, I perform a situational analysis of the participatory workshops detailed in Chapter 3. Focusing on instances where participants ask questions about AI systems, I identify the actors, the ideas, discourses, objects, relations, sites, events, etc. that are relevant. I start by creating an abstract and messy situational maps. I loosely categorize the elements to create an semi-organized map as shown in Figure 4.1. The map includes categories such as place, policing, context, existing knowledge and feelings, and lived experiences. I develop multiple iterations of such maps to understand explanation contexts.

Next, I develop multiple relational maps where I attempt to understand how different elements relations to each other. For example, how do 'policing' and 'feelings' interact to give rise to explanation needs. Examples of these maps are shown in Figure 4.2.

Situational Analysis and Situational Maps serve as an effective tool to understand the components that interact to give rise to explanation contexts in which users' explanation needs are situated. In the findings section, I describe these contexts for three of the five workshops I conducted.

4.4 Findings

In this section, I report on the explanation contexts that emerged as I mediated processes of knowing and explaining predictive systems. Out of the five workshops I conducted, I briefly describe partial contexts of Workshop, 1, 4, and 5. These workshops serve as good exemplary cases for this work.

Figure 4.1: Ordered Situational Map to visualize explanation contexts

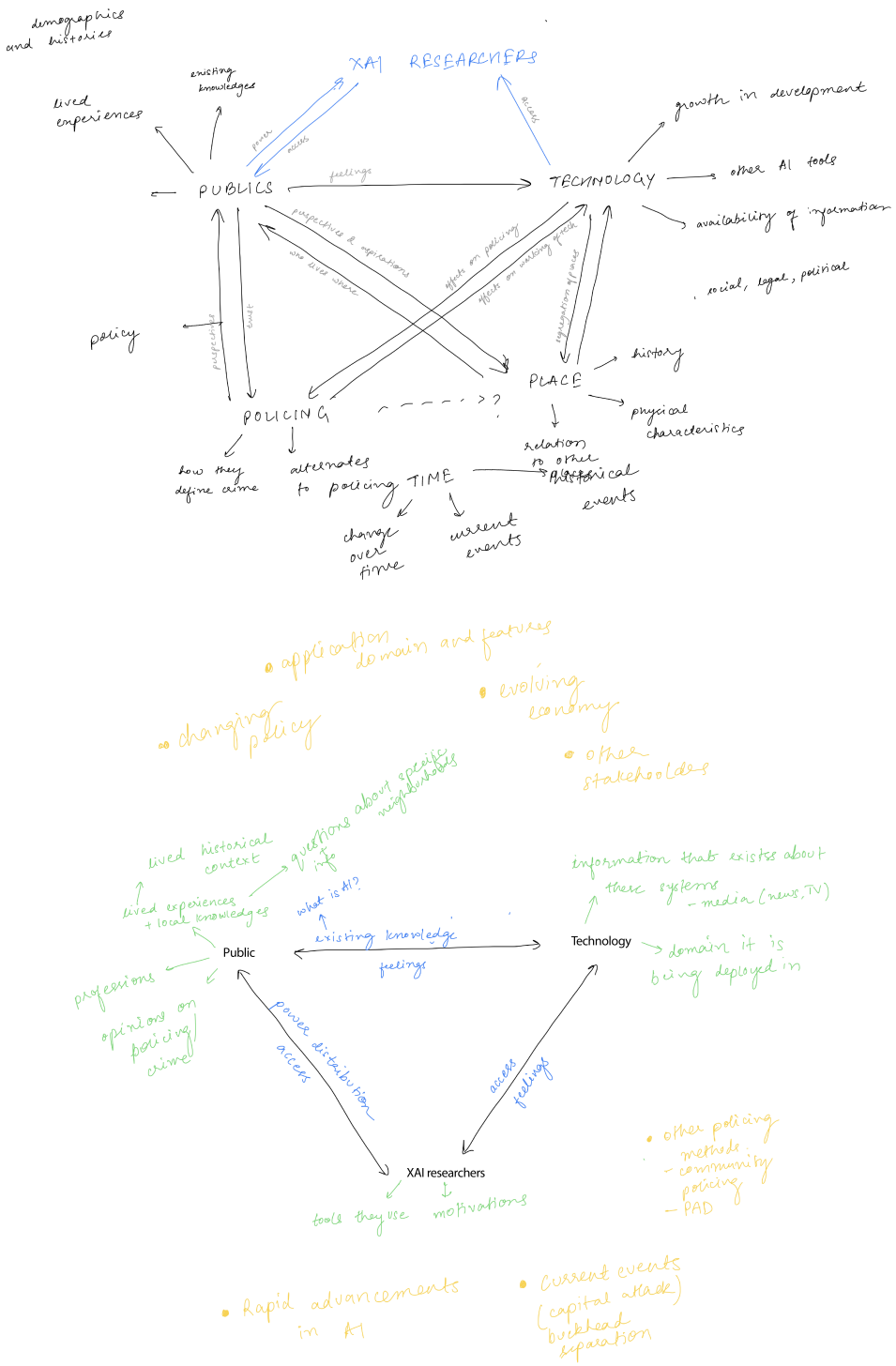

Figure 4.2: Relational maps to visualize how elements underlying explanation contexts interact with each other to give rise to users' explanation needs.

4.4.1 Workshop with police reform group (W1)

W1 invited twelve members of a police reform group that offers supportive services to those experiencing extreme poverty, problematic substance use, or mental health concerns, with the goal of reducing their arrest and incarceration rates. At the beginning of the workshop, participants expressed their feelings toward predictive policing tools calling them *"really terrifying"* and *"scary"*. Their fear was driven by *"rights issues", "reading of cases of people being exonerated from convictions with AI-based evidence", AI being "rolled out so fast", "need for compassion in policing that AI lacks", ability of AI to "decide the outcome of someone's life and future"*. These feelings of concern motivated their presence and desire to know more about the workings and effects of predictive policing tools. Their collective skepticism towards predictive tools including *"facial recognition"* or *"ads"* drove their fear of place-based predictive policing.

Owing to their position in and relationship with the institution of policing, they were highly skeptical of technology usage in policing. They firmly believed that 'crime', as defined by the law and police departments is unjust. As such, they inferred what crime would mean for predictive systems. They said:

> "[Our] definition of crime is very different than what AI systems refer to as crime. So, like drugs are not crime to us. So, this [predictive system] is a lot about like tracking drugs. But we think drugs are a public health issue and not a safety issue."

Concerned that police may categorize public health issues as crime, they asked what definition of crime is the predictive tool using. They questioned the motives of predictive systems: *"If you are going to criminalize these public health issues then it may be more beneficial to predict where there will be these public health issues"*. The intentions of the tool— to arrest people or help them— dictate the predictions a tool makes and the policing actions that follow the predictions.

Drawing on their experiences working with and living in diverse communities, they discuss the disparity in neighborhoods. The relationship of the neighborhood residents with each other and their collective relationship with the police would directly affect the workings of predictive tools. They explain:

> "What we see on a daily basis, areas like Buckhead, they give us the most referrals for criminal trespassing, and it could just be an unhoused person sitting in the park in the middle of the day. They just don't wanna see them. But on the south side, the more impoverished communities, we don't get calls like that for unhoused people because they are friends and family, they don't look at them like criminals and they support them. My mom calls the person on 11 and 23 on Cambleton road, she calls him uncle. I know his face and can point him out. If you are familiar with them, you wanna see them living. You just want them to get help."

Neighborhoods that consider outsiders to be 'threats' are more likely to report them to the police. Another participant builds on this and suggests: *"If you go off of 911 calls you have to see who calls 911. Because not everyone does. So the data is already going to be skewed."* They present a need to inquire how the data that is fed into predictive tools is collected and whose perspectives does it capture. They fear that disparate reporting patterns, if not acknowledged and addressed, can reinforce harmful policing practices.

The group emphasizes that it is essential to consider not just the data being put in and the prediction that is the output, but what we do with the predictions: *"Are they gonna start going out and arresting people or are they gonna use the fourth amendment and protect our rights once that data is gathered. I think it is about controlling the police in a way that protects our rights."* This opens another explanation need for learning about how police and other stakeholders act on the tool's prediction. They also discuss what this action means for community trust in police and holding police accountable. A *"disconnect between real events and AI predicting events will add to the distrust in a world where there is already*

a lot of distrust... how do we prevent them from terrorizing certain areas and hold them accountable for where they decide to go?" Their worries and questions are grounded in the current state of the institution of policing. They say: *"For predictive policing to have any effect of public safety, policing itself would have to have an effect on public safety... "* They ultimately ask about how this specific tool betters or worsens the current state of the institution of policing.

4.4.2 Workshop with civic funding agency (W4)

W4 brought in 11 members of an organization that leads, manages, and funds local, community-centered projects across Georgia. They considered the workshop *'timely'*. Unlike Workshop 1 described above, most of the people in this group were *'cautiously optimistic'* and wanted to understand how they could make the most of predictive technologies while managing their harms. They hoped to deploy technologies to reduce crime and increase economic opportunities in the American South.

When discussing predictive policing, participants speculated that the tool may drive police officers to areas with petty crime to increase their arrest count. An increase in arrest count may help demonstrate the effectiveness of the tool for wider adoption in the country. Given their understanding of the economic growth needed to fund projects, they expected a report of higher crime and arrest rates to be a financial motivator. However, they questioned whether an increase in arrests would truly promote public safety. This led to them asking:

> "I have a question about.. is that [arrest count] the metric police are using to say that they are efficient? I know historically they have used such a metric but recently there have been studies on how that is not the right metric and how arresting people for petty crimes is not the right metric. So, I wonder if police departments have started to shift their metrics that show they are efficient. If they have changed, then this tool may be able to help."

As such, they inquired about the methods used to assess the effectiveness of predictive

practices across police departments. They hoped that a more holistic method than 'increase in arrest count', could help prove useful to measure the impact of the tool. However, they stated *"How can you know, all these police departments are autonomous in their own ways so who knows what their objectives are?"* They discuss the challenges of opaque and decentralized police departments that may measure and report their arrest counts or safety impacts differently.

They draw parallels between crime data or 911 data with 311 data about non-emergent civic issues, such as infrastructure repair. From their work, they are aware that over-reporting of 311-related issues in certain areas draws economic investment there over other places. They ask how disparate reporting patterns may affect crime predictions.

During the workshop, participants marked high crime spots on the map based on their own perceptions of space. They revisit those markings and say *" we have so many circles here, we have little five points and waterbouys down here "*. They call the marked crimes in those areas *"very minor and visible"* such as *"drug use, which is not a crime that is actively hurting anyone"*. They were trying to make a point about how a quantitative measure of more may not always mean decreased safety for the public. They inquired about how the predictive tool measures the seriousness of crimes.

The conversation about drug use and crime led a participant to ask *"We have talked about public safety a lot. Have you looked at cities that have started to decriminalize certain things? Some of it is policy right? Historically some crimes should not be classified as crimes and should not be fed into the systems. How do these systems change with ordinances and laws? Like Houston, they decriminalized marijuana and one of their elected officials ran on the fact that arrests went down."* They asked if such policy changes resulted in data cleaning or not.

In initial introductions before the start of the workshop, the director of the group said that according to her, poverty is the main cause of crime. This was reinforced multiple times in the workshops and led to participants asking *"Are there efforts to predict why*

people commit crime instead of where? Because then maybe people need food in a certain neighborhood." "Are we getting people the right services and what do those services look like?" They asked about how predictive tools could be used to provide care services to those in need instead of arresting them using the force of police. They speculated possible responses together:

> "Crime is too big of a genre .. to think about how we can do this.. There are different types of crimes that may need different responses, maybe we can see the different types of crimes that are happening and if there are pockets of different crimes in different places and deploy different kinds of resources...[another person continues] doesn't that call for partnerships, that's when the community improvement district steps up to work with the department of public safety, that works with the civic non-profit organization, etc coming together, to address a place-based problem that is globalized to a condition."

Through these partnerships between different civic entities, we can begin to address place-based public safety concerns. They wondered how the predictive tool affects such partnerships currently as well as the potential it has in doing so in the future.

4.4.3 Workshop with educators (W5)

W5 invited a group of 8 teachers who came together as a non-profit aimed at forming meaningful relationships and collaborations between educators. Participants taught subjects such as the theory of knowledge, geography, and high school science in different schools spread out in the city of Atlanta. Many of them were participating to better equip themselves with information on civic AI systems in a world where they are suddenly being *"bombarded by AI"*. Additionally, they sought skills that would allow them to train young minds to be more critical as they grow up in the world of predictive systems.

The workshop started with a discussion about the goals of place-based predictive policing, i.e. efficient deployment of police forces. Participants talked about their desire for

a *"layered response"*. They compared policing practices in the United States with other countries and explained *"here we only have one type of cop with a gun—they are reactive, in other countries, there are cops walking around whose response is not gun"*. They also compared current policing mechanisms with historical accounts: *"I remember when we used to have sub stations... It would be nice to have police on foot who are walking around and getting to know people. Like the oldies. There are programs where police are able to get affordable housing in neighborhoods that have more crime. They used to have it where they built Olympic housing."* However, a participant explained that the response and its nature would depend on the urgency and the severity of the crime being conducted. This discussion ended with a question about the type of crimes that predictive systems currently focus on or should focus on. Participants ask: *"Are we thinking only of crime that can impact ...when we are thinking of crime, what is the definition of crime?"* As stated above, this question has repeatedly been asked in previous workshops. Deployment of police force was considered appropriate only for severe or violent crimes and therefore, participants wondered, what crimes were being predicted by the tool.

Continuing the discussion on the type of crime the predictive systems focus on predicting, they recounted recent political crimes, that may be unusual, but very impactful. They asked: *"Are they using it [predictive system] to figure out if next election we will have another charging of the capital? Where are the white supremacists? Are we looking for this?"* They considered how a predictive tool could have been useful in preventing historical crimes of a specific nature and asked if and how was this a possibility for preventing future crimes.

At numerous points during the workshop, participants drew on their own experience of space asking questions about specific neighborhoods. They wanted to know about seemingly wealthy neighborhoods that experience high crime rates. They asked *"What happens in a neighborhood like Buckhead where the income level is quite high but there seem to a lot of crimes like gun shots.. You always hear about the Publix etc. How does that inter-*

sectionality work where there is high income but lots of crime?" With this question, they wanted to know how the socio-political characteristics of a neighborhood may influence crime predictions and that means for the workings of the predictive tool.

They also asked questions about the effect of changes in neighborhoods, such as social or infrastructural growth or decline, on crime predictions. Early on in the workshop, the group discussed the relation of food deserts, lack of economic opportunities and social resources, and access to healthcare and education, with crime. They observed how different areas in Atlanta are transforming and asked: *"I would be interested in how neighborhoods have changed over the years and how has that changed the data analytics... I wanna see how adding more grocery stores change an area.. it would be interesting to see what is the impact around Old Fourth Ward since they closed the hospital.."* . By tracking the changes in neighborhoods and their affect on crime, they hoped to identify methods to disincentivize crime by meeting basic human needs.

Participants asked questions about what safety means for people in diverse neighborhoods. A participant laid out a detailed account where he compared two MARTA stations (MARTA is the public transit train system in Atlanta), East Point MARTA Station and College Park MARTA Station. He described how the College Park MARTA station has several private schools and has a large police force protecting students as they leave school. In contrast, East Point Marta Station, which is within 2 miles of College Park, has seen 2 murders in the last 4 months. Would this, they ask *"be an argument FOR predictive policing?"* Another participant adds nuance to this *"With the tri-city moms there is a double-edged sword—I want to protect my kids but who am I protecting them from? Themselves or the police?"* They questioned how the presence of police may affect the feeling of safety in diverse neighborhoods.

Above, I provide a report of the explanations contexts that emerged from meaningful questioning of AI systems. I demonstrate how publics asked questions in large part due to their feelings about, knowledges of, or experiences with predictive domains (polic-

ing), predictive subjects (spaces), predictive backdrop (current events), and predictive tool (place-based predictive policing). I discuss these relations in more detail in the following section.

4.5 Discussion

In this section, I draw on my findings to present a framework of actors and their relations with predictive systems. I argue that public-centered explanation needs emerge out of these relations of feeling, knowing, and experiencing broader aspects of the predictive system such as the prediction domain, prediction subject, predictive tool, and predictive backdrop. Below, I explain these relations by drawing on workshops. Next, I discuss how this framework, as shown in figure 4.3, can be used to advance human-centered XAI research.

4.5.1 Framework

Publics and Prediction Domain

In this work, I identify the prediction domain as 'public safety', or more specifically 'policing'. People's beliefs, familiarity, and experiences with policing as well as other methods to promote public safety led them to ask about the goals of the predictive system, the impacts it may have, and who will bear the burden of harmful effects. They also considered the history of policing, the tools and standards used to measure effectiveness, and if and how they translate to predictive tools today. They drew contrasts with other policing methods in different places and in different times to investigate how predictive tools intervene in existing practices— do they reorient them or do they reinforce them?

The starkest example of people feeling strongly towards 'policing' was observed in Workshop 1 conducted with members of a police reform group. Their professional work put them in a unique position to witness the effects of policing. They possessed detailed knowledge of how police forces are incentivized to function and thereby questioned the

position a predictive tool may take in this domain. Lastly, many of them identified as black people, born and raised in Atlanta, and had themselves experienced disparate treatment by residents and police alike in different neighborhoods. Ultimately their intricate and grounded knowledge of place-based public safety needs and conditions guided their investigation of the workings and effects of place-based predictive policing tools in Atlanta. Some recent work has also highlighted that people draw on their knowledge of police brutality along with their own experiences of racial profiling to demand oversight over the data that feeds the algorithmic models [190].

Publics and Prediction Subject

In this chapter, I identify prediction subjects as 'spaces or neighborhoods' in our cities that are categorized by predictive tools. People's perceptions and experiences of the places they live and work in led them to ask place-specific questions about predictive tools many times naming the areas such as 'Buckhead' or 'Old Fourth Ward' and asking about the quality of predictions made there. They discussed the socio-political characteristics of neighborhoods such as mean income, or the resources present in a neighborhood such as healthcare, schooling, or grocery stores, and asked to know how these elements affect, and are affected by predictions.

The prediction subject consists of not just the *physical* spaces that are subject to predictions. The spaces include the lives of the people who live there, their histories, political leaning, relationships with each other, reputation, and so much more. Participants reflected on their relations with the people in their neighborhoods, contrasted them with other neighborhoods, and asked if and how communities that see homeless people or outsiders as threats report data and if that would affect the predictions. Predictions may not just be impacted by, but also impact people in different neighborhoods differently based on their relationship with policing. Participants asked what these disparate effects may be.

Publics and Predictive Backdrop

In this work, I identify predictive backdrops as 'local and global' events surrounding the use of predictive tools. Several participants were motivated to participate due to their knowledge and experience of the booming AI industry and its harmful effects. They wanted to make sense of the fast growth of AI they felt around them. Participants constantly drew on their knowledge of relevant policies and current events to ask questions about predictive tools. They focused on specific instances such as the Capitol attack on January 6th 2021, where over two thousand people violently invaded the United States Capitol Building in Washington, D.C., in support of then-U.S. president Donald Trump, two months after his defeat in the 2020 presidential election [191]. They asked if and how predictive tools may help prevent such events. They asked about the effects of constitutional rights such as the Fourth Amendment that protects people against unreasonable searches and seizures and changes in laws such as decriminalizing marijuana, on the use of predictive tools.

Participants also inquired about the need for cooperation and partnerships between different governmental and non-profit organizations required to deliver on the promises of safety that these tools made. They considered the reporting of other civic improvement data that they were more familiar with, such as the 311 data, and compared that with 911 data to ask about the effects of disparate reporting patterns and effects.

Publics and Predictive Tools

For this paper, I identify prediction tool as 'place-based predictive policing models' deployed to predict crime hotspots in a city. While some groups were wary of the use of predicting tools in policing and wanted to know about how harms are being managed, others were cautiously optimistic and wanted to know how these tools can be effectively used to prevent crimes. The feelings participants had towards the predictive tool influenced their explanation needs and goals.

Participants also asked questions about place-based predictive policing in relation to the

broader world of AI including other tools used for public safety such as facial recognition systems or fingerprint scanners, as well as beyond public safety such as 'ads', 'generative AI', etc. Due to their more direct engagement with other automated tools, their questions were grounded in drawing contrasts with those tools. Oftentimes, participants' broader perception of AI (dystopian, privacy invading, optimistic, the next big thing), rather than their perception towards a specific tool, informed their questioning in the workshops.

4.5.2 Implications

Exiting work on human-centered XAI has majorly focused on identifying user needs based on their technical knowledge and functional roles. The framework presented in this chapter advances existing work in two primary ways. One, I find that explanation needs emerge not just from user knowledges, but also from their feelings and experiences. Together I call knowing, feeling, and experiencing— *relating*. Two, I demonstrate that publics *relate* not just to the predictive tool, but also to the prediction subject, context, and backdrop, to seek explanations about the workings and effects of predictive systems.

For us, *knowledge* refers to information people gain through formal, practical, or social means; *experiences* refer to direct interactions or lived experiences; and *feelings* refer to the perceptions people may have. These *relations* are entangled in many ways. For example, the knowledges and experiences of people inform their feelings and their feelings and experiences may inform the knowledge they seek and gain. Yet, the distinction between these three relations helps XAI researchers detangle them to an extent that they can be considered distinctly important, in their own right, as I design explanations for publics. Such a distinction would help researchers avoid reductive assumptions such as presuming that if a person is knowledgeable about AI, they may only need explanations for technical debugging. The person may seek explanations based on their lived experiences, such as justifying the increased police presence in their neighborhood, or to understand why people are worried about the use of AI in policing.

The *relations* defined above are not only with the predictive tool (the machine learning model that makes predictions about crime hotspots). In addition to predictive technologies, publics relate to the prediction domain (policing), subject (space), and backdrop (local and global environments). Once again these elements are entangled. For example, policing practices may differ with spaces. The tools may not even be used in specific spaces. The backdrop may include historical policing practices in specific local spaces. This distinction can help XAI researchers consider how publics relate not just to the predictive tool alone, but also to these surrounding elements. People's knowledge of discriminatory police practices may lead them to ask about the protocols police forces follow when acting on a prediction. Their experience of space, such as observing gentrification, may lead to ask about the role of increased police presence in the same. These examples I provide above are inspired by the workshops.

These relations with the predictive elements that I define in the paragraphs above are not exhaustive. Instead, they are a starting point into considering how human-centered explanations can account for the situated and grounded explanation needs of diverse publics. Some other researchers have also mentioned such relations with predictive elements. For example, Suresh et al. discuss personal knowledges of the milieu that aligns with my conceptualization of *experiencing* predictive elements. I echo their findings and add nuance to their work through this framework by (1) providing empirical evidence by reflecting on real-world contexts of explaining and knowing predictive systems, and (2) detailing the relations with predictive elements in ways that can be useful for XAI researchers in identifying users' situated explanation needs.

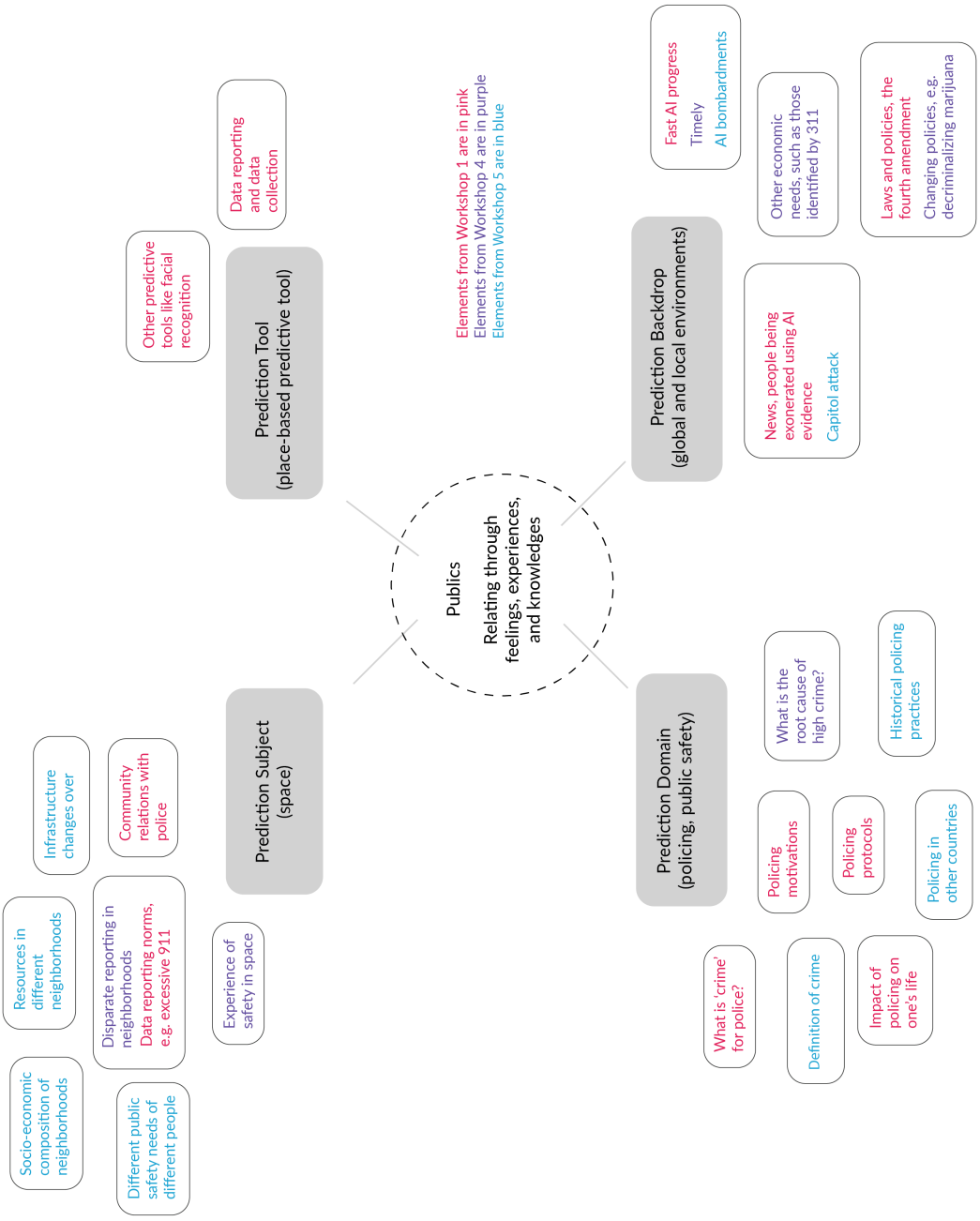

Figure 4.3: Framework to understand explanation contexts

CHAPTER 5
SYSTEMIC EXPLANATIONS BY LOCAL PUBLICS

In the last chapter, I discussed how explanation needs are situated in diverse explanation contexts. These explanation needs inquired not just about the predictive tool but also the complex systems that such tools become a part of. As such, this chapter asks: *How can we move towards a better understanding of the expansive and complex socio-technical assemblages underlying AI systems?*

AI systems emerge out of broad networks of materials, relations, cultures, institutions, and histories that may affect societies in unjust and harmful ways. In this chapter, I argue that local publics can partially explain how algorithms interact with society to affect local contexts. I report on our efforts to engage with these diverse partial explanations through the participatory mapping workshops detailed in Chapter 3. I find that partial explaining can (1) collectively advance our understandings of AI systems as socio-technical assemblages, opening up novel critical questions about these systems, and (2) identify gaps in current algorithmic explanations thereby creating explanation needs that we can attempt to collectively address. I call for the creation of spaces and systems where such engagements of partial explaining and knowing can happen.

5.1 Introduction

Today, our efforts to explain AI systems are limited to opening the 'black box' and learning about the data, variables, algorithms, and models that it builds on. As models become more complex, are constantly changing and evolving, and are contingent on countless unknown data and their categorizations, this black box becomes tougher to look into [178, 192].

Additionally, existing pursuits towards transparency are overwhelmingly technical and disregard how algorithms interact with broader networks of materials, relations, cultures,

institutions, and histories to affect societies in unjust and harmful ways. A meaningful examination of an AI system requires us to engage with the socio-technical assemblages in which it is placed [193]. For example, understanding predictive policing systems as socio-technical assemblages would include learning about the spaces they categorize as crime hotspots and their characteristics, their effect on relations between police and communities, and the policies, laws, and protocols they engage with. In what follows, I use the term 'AI systems' to refer to these broad and complex socio-technical assemblages.

There is a need to (1) overcome the epistemic barriers presented by opaque 'black-boxed' algorithms, and (2) situate algorithmic systems in broader networks of spaces and environments that affect and are affected by the algorithm [97, 10]. I follow the lead of researchers who address these challenges by proposing to 'look across' the black box, rather than 'look inside' them [141, 178]. Understanding algorithms as part of wider urban assemblages that continuously modify and evolve can surpass the dependence of knowing AI systems merely by opening the black box and would allow us to understand AI systems more holistically in relation to the socio-technical networks they become a part of. To that end, in this chapter, I ask— How can we move towards a better understanding of the expansive and complex networks underlying AI systems?

In their entirety, AI systems always remain unknowable [194]. As with all knowledge, they can only ever be known from specific standpoints and in specific settings [195]. Publics, from their own standpoints, are experts of their own domains [196], and possess knowledges that can add to, complement, or even challenge the perspectives of experts [178]. I empirically explore the role local publics can play in explaining elements of the broader AI system from their unique and grounded standpoints.

To do so, I draw on the participatory mapping workshops I conducted that mediate collaborative processes of explaining and questioning civic AI systems. As detailed in previous chapters, these workshops engaged with diverse publics including police reform groups, urban planners, neighborhood leaders, funding agencies, and teachers. In the pro-

cess of conducting these workshops, I find that local publics are positioned well to partially explain how algorithms interact with society to affect local contexts. Locals directly engage with components of AI systems outside of the 'black box' and can make known, amongst other aspects, the environments AI tools are deployed in, the cultures and norms they invade, and the lived experiences of the problems they attempt to address. As such, local publics possess partial explanations of AI systems that can come together to raise meaningful and grounded questions about the effects of AI systems.

Ultimately, I argue that bringing partial local explanations together can help us collectively advance our understanding of AI systems as socio-technical assemblages opening up novel critical questions, and help identify gaps in current explanations thereby creating explanation needs that we can attempt to address. I call for the Explainable AI (XAI) community to design spaces that can invite local publics to collectively uncover nuanced networks underlying complex AI systems.

In what follows, I summarize existing approaches that explain AI systems as socio-technical assemblages and if and how publics have been invited to participate in the design and governance of AI systems in the 'background' section. Next, I briefly summarize the data analysis methods for this chapter in the 'data analysis' section. I report on the explanations publics provided in the 'findings' sections. I end by discussing the role of XAI researchers in designing systems for partial explaining and knowing.

5.2 Background

5.2.1 Exiting methods to explain AI systems as socio-technical assemblages

Majority of existing efforts to explain AI focus on technical components such as algorithmic features and their weights, training datasets, source code, and performance metrics [99]. Recently, the XAI community has called for moving beyond technical transparency and considering other socio-technical elements impacting AI decision-making. Few but growing frameworks and concepts are being proposed. Ehsan et al. introduce social trans-

parency, which includes making transparent the technological context, decision-making context, and organizational context surrounding algorithms [64]. Similarly, Kroeger et al. account for user literacy and propose 'social explainability' as a concept that explains not just the AI system but the AI ecosystem [197]. They argue that insight into the institutional systems that AI becomes a part of can help users trust the institutions deploying AI even if they do not understand how an AI tool works [197]. Cobbe et al. argue that current XAI methods provide insufficient information to regulate AI systems and propose 'reviewability' as a framework for more meaningful transparency across the entire AI lifecycle [198]. Geiger et al. use a "sociotechnical systems approach" and propose reimagining 'audits' as tools that don't settle 'matters of fact' but instead that open up investigation into 'matters of concern' [199]. Such a change in perspective, they urge, will prompt us to ask questions about algorithmic systems in relation to the organizations that deploy them. Yu-Shan Tseng specifically focuses on Urban AI and proposes 'assemblage thinking' as a methodology to study the relations between algorithms and the cities they are situated in [97]. They present a case study of vTaiwan, an open-source algorithmic platform, as situated in complex urban assemblages. They describe in detail its political and collaborative origins and how it creates communities for democratic engagement.

Community-centered methods have also been proposed or employed to study AI systems as socio-technical assemblages. Rob Kitchen echoes the need to study algorithms as situated in broader socio-technical assemblages. He recommends conducting interviews and ethnographies of particular AI systems to shed light on the principles they adhere to [193]. Seaver proposes using ethnographic methods, living with-in algorithmic systems, to know not just how algorithms work, but how they come to be [192]. Stop LAPD Spying Coalition worked with an activist organization called the Free Radicals and investigated the use of Predpol in Skid Row, a community in downtown Los Angeles. They proposed a framework called the 'Algorithmic Ecology' that maps, visualizes, and communicates the relationships of power that surround any algorithmic technology [200].

Researchers and artists have attempted to visualize the socio-technical nature of AI systems for better user understanding. Kate Crawford and Vladan Joler designed a visualization titled 'Anatomy of AI' that traces AI throughout its lifecycle and makes transparent the resources that help develop an AI system [201].

These works motivate the need to understand AI systems as complex and entangled networks of social, historical, cultural, and technical components. I add to this rich foundation of methods to explain AI as socio-technical assemblages.

5.2.2 Explaining AI systems by local publics

Another shift in XAI efforts has been guided by the desire to move beyond merely 'experts' providing explanations to 'non-experts'. Experts, from their own standpoints, are positioned well to explain specific parts of AI systems. However, scholars are increasingly recognizing the need to engage with the public to understand the effects of algorithms. Eyert and Lopez's 'transparency as a communicative constellation' calls for multi-directional transparency where tech experts don't just teach, but listen and learn from publics [86]. In a similar vein, Nicenboim et al. call for co-creating an understanding of AI systems with both users and artificial agents [178]. They argue that explanations and methods of understanding offered by citizens are no less legitimate than the explanations offered by experts. Engaging with external actors and diverse publics can help identify relevant algorithmic concerns and needs [178]. Sloane echoes these values and presents participation by a more diverse set of actors, especially in ways that are "unscripted, unpredictable, and are beyond the grasps of bureaucratic control", as a technique to 'unblackbox' AI systems [202]. Corbett and Denton join fellow researchers and respond to the technocentric turn that the XAI communities have taken, by making calls for 'bringing people to transparency' [65]. 'Participatory AI', as defined broadly, provides various degrees of power [203] to the public to inform the design, use, assessment [204], advocacy [115], and governance [205] of AI systems.

Growing, but still nascent work, has involved users as experts who can explain AI systems. Barnett and Diakopoulos propose crowdsourcing as an effective method to anticipate societal algorithmic harms [206]. They utilize the diversity of Amazon Mechanical Turk workers to describe the impact domains of algorithms used by the U.S. government. Marian presents a brilliant case study where parent groups and researchers come together to understand the (New York City) NYC School Matching Algorithm [207]. The authors crowdsourced information about the lottery numbers that students received and the schools they were matched to understand the role lottery numbers play in the matching algorithm and identify cut-offs for various schools. Their results were made available publicly to help keep student applicants in the following academic years informed about how the system works. Other 'citizen-science' projects have attempted to gather distributed information about technological systems. The 'ban the scan' project by Amnesty International mobilized volunteers to share information about the presence of surveillance cameras on the streets of NYC and Hyderabad [208]. By doing so, they were able to explain the omnipresence of discriminatory facial recognition systems and call for a policies that ban these tools. These examples of crowd sourcing are inspiring starting points. However, they are still driven by researchers and not publics. This does not allow publics to draw on their multi-faceted expertise. Rather they provide information as requested by the researchers. They involve publics only to the extent that they can provide experiential data points without considering them experts of their own elements in the broader AI system.

I build on these works described above in three ways: (1) I provide publics with an open platform to explain distributed parts of an AI systems based on their expertise, (2) I bring the expertise of diverse publics together instead of focusing on one aspect explained by one stakeholder type and , (3) I focus on nuanced grounded explanations instead of factual data points.

5.3 Data Analysis

This chapter aims to investigate if and how local publics can partially explain elements of AI systems in the capacities in which they interact with these systems. To do so, I start with identifying and pulling out instances of partial explaining and questioning from the workshops to form a repository. Next, I write memos [122] reflecting on if and how this explanation can help us better understand AI systems as socio-technical assemblages alongside novel questions that it opens up.

This was followed by a reflexive thematic analysis [209] where I categorized partial explanations offered by publics into elements of AI systems that they explained such as—spatial, social, relational, historical, policy, lived, etc. Through further refinement of these themes I identified the categories of explanations that organize the findings section below. These categories are not exhaustive or mutually exclusive but serve as an example of the variety of partial explanations users can provide about socio-technical AI systems.

5.4 Findings

In the workshops, diverse local publics partially explained components of AI systems such as: (1) spatial, social, political, and historical characteristics of the places that are subject to predictions, (2) institutions, cultures, and contexts that algorithms engage with, and (3) lived experiences of public safety and related resource allocations including care, financial, health, food, and shelter resources. These explanations revealed novel features of the broad and complex AI system, prompting critical inquiry into how these features could impact society.

5.4.1 Experience of safety in diverse spaces

Participants explained the on-the-ground reality of spaces that are subject to predictions. In workshop 2, the group was talking about the places they think are most in need of public

safety aid. One of the participants talked about a specific neighborhood where he goes to work every day. He explains:

> This is where our office is and it is in Mercy Care which is like a free clinic so 15, 16 thousand mostly unhoused uninsured patients a year. And there is this bridge on Hillyard street and there are people who sleep under that bridge. So I was thinking of the vulnerabilities of the people who live around it so they can access medical services but don't have housing. They are exposed to all kinds of things.. It's also gentrifying, so they are getting displaced further away from where the services are. We have pretty much criminalized homelessness, so they are maybe at risk of police interactions or something. [W2]

As he described this area, he highlighted the importance of considering whose safety are predictive tools prioritizing. With a better understanding of the realities of spaces subject to predictions, we can assess how the deployment of more police forces may affect the diversity of people that inhabit the spaces. He also talked about the socio-economic characteristics of the neighborhood— rising gentrification and increasing distance from health care services and how that affects the safety of people living there. Lastly, he explains the social and political tendency to criminalize homelessness. These tendencies and corresponding policies dictate what counts as a 'crime' and who is detained by the police deployed by a predictive tool. In another workshop (W3), participants explained the effects of the use of a predictive tool in a specific neighborhood. One of the explains:

> I can tell you that a predictive factor in Mechanicsville will not be welcome. No one wants to be overpoliced. Mechanicsville, East Atlanta, where gentrification is happening, people are being pushed out of their communities and then they are being heavily policed. That doesn't feel safe, that feels unsafe. Something to think about is that how does it land on people who have been disinvested in and marginalized. These tools don't tell the whole story. They

> say they want to address safety issues in this bubble, that is law enforcement and not system changes. [W3]

The participant directly explained how the use of the tool will affect specific communities and if and how it may work to reinforce ongoing injustices. She provides a clear explanation of the oppressive contexts in which the tools will be deployed and calls for addressing issues not by law enforcement that will further marginalize communities but by systemic change.

5.4.2 Norms and cultures in space

In workshop 1, a participant that works for a police reform group describes her interactions with reports for criminal trespassing in upper-class wealthy neighborhoods of Buckhead. She explains:

> What we see on a daily basis, areas like Buckhead, they give us the most referrals for criminal trespassing, and it could just be an unhoused person sitting in the park in the middle of the day. They just don't wanna see them. But on the south side, the more impoverished communities, we don't get calls like that for unhoused people because they are friends and family, they don't look at them like criminals and they support them. [W1]

Overreporting often invites more police presence resulting in unwanted violent encounters between police forces and marginalized communities. Sometimes, as another participant in workshop 3 explains, community members are encouraged to call 911 repeatedly so more police forces are deployed in their areas:

> And the police in neighborhoods are not shy of telling people you need to call 911 to get that data point because that helps direct the zone commanders to localize resources to those areas. So, some people, maybe not excessively, but

> some folks are .. they would become hotspots where it has become the social norm to call 911 in comparison to other areas where thats not something..[W3]

The participant continues to explain how some neighborhoods have special committees designed specifically to address crime by requesting more police presence and surveillance cameras or even self-policing their neighborhoods by registering retired police officers as volunteers. On the other hand, some neighborhoods prevent police encounters as much as possible. Such norms and cultures in different neighborhoods are reflected in place-based data that inform the predictions made by algorithmic tools.

5.4.3 Data Contexts

One of the participants in Workshop 3 leads a violence prevention program affiliated with one of the largest hospitals in the city of Atlanta. As part of the program, they analyze data related to firearm injuries, while working with community partners to reduce the risk of recurrent injury. She describes the data her team collects and explains:

> And if we are talking about violent crime, one of the things that I wanna say is that 80 % of our cases are not reported to law enforcement.. So there is huge discrepancy in what we see at the hospital and what we know ..what APD (Atlanta Police Department) or Dekalb county's data reflects.. [W3]

Through her work, she has access to data that explains the limits of existing crime-related data. Many people who come to her hospital to seek care do not officially report the crimes that they become victims of with the police department. Their reasons may vary but it foregrounds the sheer amount of data that is missing from the training data.

5.4.4 Explaining spatial characteristics

'Five points' was an area repeatedly identified as a crime hotspot across workshops. In workshop 2, which brought together a group of urban planners with an understanding of

urban space and its effects, one participant explained why five points may be seeing a high crime rate:

> I have seen people get ticketed for jay walking at five points. I am not sure anywhere would ever get ticketed for that anywhere else and it is like they were pretty specifically targeted. It is like they are trying to keep people out of certain spaces. And to catch a bus, you have to jay walk, there is no crosswalk.

As he described targeting people for petty crimes in five points, he also highlighted how the spatial features of the area make it easier for police to ticket people over crimes like jaywalking. Later in the workshop, a participant described another such intersection where the spatial features promote an increased police presence:

> Another thing that is very unique right here [highly policed area] is that there is cross-jurisdictional cooperation between 13 different police departments. There are 13 different kinds of police that can stop you, arrest you.. It's a residual policy from the Olympics. It is a huge reason why a lot of these spaces are so policed because they are gigantic mechanisms for moving people in and out for events and protecting their cars essentially when they are inside.

He explained how historical events, such as the 1996 Olympics held in Atlanta, informed the design and segregation of spaces in lasting ways. Such segregation, that were designed for alternate purposes and contexts, continues to affect the people who live in those spaces today.

5.4.5 Explaining policing institutions

Participants in Workshop 1 worked as part of a police reform group. They offered care and support services to people experiencing extreme poverty, problematic substance use, or mental health concerns to reduce arrest rates. They work in collaboration with police

departments who may sometimes divert 911 calls to their organization. Participants in this organization were driven by the cause they support, seeking police reform. They were knowledgeable about policing works and its many shortcomings and were able to explain the institution of policing in relevance to the predictive tool. One participant explains:

> I will say that it actually started as a way of slave control. Then it became something that helps the white class in the reconstruction era to control free black people. And so it evolved from that. Until there is true reform in how these traditions of policing, then you can't bring something like this in the picture. [W1]

They explain the historical origins of policing that are still prevalent in how police operate today. With that explanation, they help us understand what is needed to promote public safety in ways that challenge the historical underpinning of the institution and policing and make way for more just practices. They also explain the current state of prisons, specifically in Georgia:

> Georgia has the most private for-profit prisons of any state in the entire country. And there are a lot of large corporations that use that free labor. [W1]

This explanation helps us understand and question the underlying incentives of predictive tools. Such a financial incentive can push large corporations to influence the workings of predictive systems in ways that drive more people into prisons for increased profits.

5.4.6 Explaining desirable futures

In workshop 4, a participant explained how civic organizations can work with the city to offer care when they observe a pattern of recurring crime:

> Downtown in Woodridge Park, which is right here [shows on map], has a high homeless population, they tend to stay in the park all day, there are a lot of

> businesses in the area who hand out food. Woodridge park recognized that this is a different kind of 'crime' so the park worked with one of the agencies downtown to have a resource officer staffed as an employee right outside the park instead of a police officer who works with the people in the area to adjust their needs and to get them out of the streets and into support services, so that is an example of recognizing a pattern of a certain kind of crime in an area and addressing it through a tailored approach.

Such real-world explanations of how we can address certain kinds of patterns and predictions with an approach that centers the needs of the marginalized can help us understand the kinds of impact predictive tools could aim to have. Such explanations can guide the redesign of protocols following a prediction to involve the use of civic resources, eventually leading to systemic changes.

These instances described above demonstrate how diverse publics may be well-equipped to explain parts of an AI systems that they engage with.

5.5 Discussion

AI systems exist as socio-technical assemblages of histories, cultures, institutions, and practices. AI tools affect and are affected by interactions with these broader networks. How then can we understand these networks and their underlying mechanisms to identify, assess, and regulate AI related social harms? This chapter contributes to existing literature on studying AI systems as socio-technical assemblages by empirically investigating the role that actors inhabiting parts of an AI system can play in partially explaining them. I find that partial explaining by diverse publics can collectively advance our understandings of AI as socio-technical assemblages opening up novel critical questions about them, and identify gaps in current explanations thereby creating explanation needs that we can attempt to collectively address.

Recently, scholars are suggesting the possibility of involving publics in explaining AI

systems. However, there has been very limited research that has attempted to do so, or has studied how to do so. Some research has shown the potential of crowdsourced data via citizen science projects in explaining parts of an AI system. This limited work has still been highly impactful in calling for more transparency, mobilizing for regulation, and identifying AI related impacts [207, 208]. Through this work, I demonstrate how these practices of explaining by publics can grow and the role XAI researchers can play in doing that.

Firstly, in this work, I followed the lead of publics' when generating explanations. I provided them with an open platform to draw on their unique and diverse expertise and explain parts of the system they have observed, experienced, or engaged with. This, as I note in my findings, generated explanations about elements I had not considered in the workshop design as well pluralistic conceptualizations of civic futures [210]. Such explanations highlighted parts of the AI system that actively shape how publics' perceive, relate to, or experience AI. These parts can now be further explored and researched to understand how they work to shape AI systems and their effects.

Secondly, I encouraged publics to explain AI systems in grounded and nuanced ways. I designed spaces and tools to help them reflect on their expertise as both citizens and workers in society to identify and explain parts of an AI system. Such grounded explanations helped us understand the relation between the technical and the social, political, and historical elements of AI, avoiding the need for abstraction or assumption. The nuance in the explanations I report are absent in existing XAI processes, keeping them disconnected from on-the-ground lived realities of civic AI.

What role can the XAI community play in supporting the generation of partial explanations by local publics? Below, I reflect on my research method along with supporting examples to describe in more detail the role XAI researchers can play in co-creating effective partial explanations with local publics:

5.5.1 Gather Partial Local Explanations

As I find in this chapter, local publics have the ability to explain parts of an AI systems through their role as actors in this system. I begin by urging the XAI community to gather such partial explanations instead of relying solely on tech experts or insiders to generate explanations. This will not only help overcome the epistemic challenges presented by black box algorithms or the access challenges due to lack of cooperation by technology makers, it would also help us understand how AI systems work as socio-technical assemblages that interact with, affect, and can possibly harm real world spaces. In seeking such partial explanations, there remain important factors to consider including: (1) who is involved in partially explaining AI systems and how their knowledges are privileged or discounted, and (2) how can tools, methods, and protocols support publics in trusting their expertise and explaining AI systems as they relate to them. In this work, I gathered partial explanations from diverse groups who I believed had a stake in the working of predictive policing. My efforts were limited by (1) who has the social capital to organize in ways that make them approachable, and (2) my ability to form relations with social groups. I used participatory methods, grounded in the places people know and consider their own to encourage participants in sharing their knowledges.

5.5.2 Organize Partial Local Explanations

Next, I highlight the need to organize and formalize these partial explanations. While partial explanations in themselves help in the development of nuanced understandings of AI, when such explanations come together, they can help identify patterns and draw insights into how an AI model works in relation to other socio-technical components more generally and related potential harms. However, as Loukissas, reminds us, 'all data are local'. So are all explanations. Like data, explanations too are created by specific people, using specific tools, and for specific goals. The challenge is to acknowledge and preserve the locality of explanations even as we attempt to organize them in effective ways. In this chapter, I

attempted to organize explanations (1) locally, by visualizing effects on a shared map, and (2) across workshops with the organizers acting as threads bringing explanations from one workshop to the other and finally together in this writing. There is more work to be done. I plan to organize the collection of explanations spatially through the use of an interactive map to eventually identify how predictive policing may affect spaces disparately.

5.5.3 Continuous Partial Local Explanations

Lastly, I discuss the need to consider how to place these explanations in systems that allow continued development and iteration of explanations. The development and organization of partial explanations are not a one-time effort. In the ever-changing world of AI, the explanations publics choose to, or have the ability to offer will also continue to evolve. Such developments will require redevelopment and reorganization of partial explanations. To promote continuous development of explanations, these processes of explaining need to be placed in systems that bring people together, time and again, around growing capacities of AI, to engage in continued and long-lasting explaining . Unfortunately, the workshops were not part of an official or robust system that had the capacity to be long-lasting. Yet, the work, done as part of a research project, with a fixed timeline and funds, was able to form a small system of five workshops over a period of a year for the continuous development and organization of partial explanations. In the process, I identified other sites that may be useful for continued engagement such as the (Neighborhood Planning Unit) NPU University [211] that invites people to learn about civic processes every spring and fall.

5.5.4 Limits

I acknowledge that such partial explanations may not always be concrete, complete, or accurate, nor do they need to be, in order to support critical engagement with AI systems. In saying so, I am following the lead of Gabrys et al. [212] who argue that citizen data may not be complete or accurate but is 'just good enough' to create a shared space for

discussion. Similarly, partial explanations offered by local publics are 'just good enough' to present concerns and launch an inquiry into the effects of AI on society.

Ultimately, I call for the broader HCI and XAI communities to design spaces for partial explaining. I elaborate on this in the next chapter. I hope that such spaces can provide systemic and long-lasting ways to support democratic and critical engagements of local publics with AI systems and their underlying power structures [199]. I would like to note that in this chapter, I do not play the role of XAI developers trying to partially explain place-based predictive policing. Instead, I serve as a design researcher identifying the role of local publics in the design of AI explanations. My goal then is not to explain, but to study explanations.

CHAPTER 6

SLOW AND PARTIAL EXPLAIN*ING*

Early in this dissertation, I started on a quest to conceptualize 'good enough explanations' of algorithmic systems. Good enough explanations of predictive systems, I argued, were not complete or objective, but were good enough to support publics in critically engaging with civic predictive systems.

We learned that explanations are good enough for *someone* in their grounded contexts. Such contexts, as I demonstrated consist of how people know, experience, and feel about predictive tools, domains, subjects, and backdrops. Next, I learned that tech experts are not the only ones who can offer good enough explanations. Publics, who interact with the socio-technical assemblages underlying AI systems from their positions in this system can explain parts of an AI system. I proposed that XAI researchers could act as moderators who gather, organize, and continuously develop a growing collection of partial explanations. Such moderation however is not an easy task.

Current work conceptualizes the 'explanation' of algorithmic systems as a noun— static artifacts that can be exchanged between groups of people without changing form or meaning. However, any transfer of information requires a system of institutions, interfaces, and methods through which it travels. As it travels, it changes meaning or form. Oftentimes, in this process, it is recreated for diverse contexts creating new questions and knowledge gaps. Explanations, through this lens, are a verb, a process—explain*ing*. *How can XAI researchers moderate these processes of explaining with diverse audiences in partial and continuous ways?*

Explaining, understood as a process, requires us to consider the (1) sites where information is generated and consumed, (2) media through which they are explained and understood, and (3) interactions that mediate the processes of questioning and knowing.

These considerations inform the design of situated and systemic processes of explaining.

In this chapter, I foreground three related design decisions that I believe designers will have to grapple with as they operationalize the practice of good enough explaining: (1) designing explaining sites (2) designing explaining modalities, and (3) designing explaining interactions. Below, I reflect on my own experience of addressing these design decisions, along with challenges, lessons learned, recommendations, and related work.

6.1 Designing explaining sites

Where does explaining happen? Existing research rarely engages with the *place* of explaining. When designing systems of explaining, XAI researchers must actively consider their sites, or their entry points [213], into explaining. The design and selection of sites can determine who gets to participate, in what capacity, and to what ends [214].

Typically, sites for explaining algorithms are the same platforms on which a user encounters the AI system. For example, the 'why ad' button on an ad will explain why a user may be seeing an ad. Sometimes, the platform provides links to other sites where a user can get a more detailed explanation of how an algorithm works along with options to control them, as with the Ad centers [215]. However, civic AI systems are not packaged in a platform that is easily accessible by the publics. *Where* then can we explain civic AI systems to the public? Below, I report on my experiences engaging with this question, the constraints I encountered, and the lessons I learned.

6.1.1 Our approach, learning, and challenges

Our approach involved exploring different sites and empirically investigating the opportunities they present along with constraints and limits. The sites I identified for initial exploration included:

1. Existing civic sites: NPU University [211] and Citizen's Police Academy [216]

2. One-on-one partnerships: Civic organizations and non-profits (W1, 2, 5)

3. Existing and scheduled training modules: Organization retreats (W4)

4. Open call workshops [217] (W3)

5. Neighborhood meetings

6. Neighborhood meet-ups

All these sites varied in what they offered for the purposes of this dissertation. Existing civic sites, like the NPU (Neighborhood Planning Unit) University [211] and Citizen's Police Academy [216] offered civic education to Atlanta residents. They were the most robust channels ensuring that I would be able to slowly, but *continuously* engage with diverse publics. They also were formalized in trustworthy ways by being affiliated with the city of Atlanta and the city police department respectively, inviting increased participation from citizens. Yet, I was unable to organize workshops at these sites because of their rigid structures involving time intensive bureaucratic processes for approvals. I did receive positive feedback from them and they encouraged us to offer courses and workshops as part of their programs. I had to forego these options due to the time and logistic constraints of this PhD. One-on-one partnerships allowed us to seek out people who I considered relevant for this work. I was also able to select groups that may benefit from learning about civic predictive systems or will be able to explain parts of these systems themselves. W1, W2, and W5 were conducted through one-on-one partnerships. Open call workshops were hosted on the Georgia Tech campus and required extensive marketing to spread the word such as via flyers and emails (see Appendix A). These sites, however, cherry-picked people who had the interest and ability to travel to the Georgia Tech main campus. Many participants interested in open-call workshops preferred virtual meetings and eventually dropped out. Neighborhood committee meetings and Neighborhood residential meet-ups were another site I explored. These meet-ups involved voluntary participation and, despite the interest in

this work, people rarely had the time or energy to participate at such sites. As such, I was unable to organize workshops at these sites.

6.1.2 Opportunities and challenges

In selecting sites of engagement, I had the opportunity to seek out publics, focusing on stakeholders that have rarely, if ever, been engaged in conversations around civic AI systems. With this opportunity, I also carried the responsibility of selecting groups for engagement. I had the power to engage certain groups who may have had the capital or resources to organize in ways that made them approachable. I was the lead moderator of all workshops. As such, I was able to weave the workshops together, sharing discussions from one workshop in the next. This created a bigger more connected site and I had the opportunity to support indirect conversations and partnerships. Unlike workshops with organizations where team members generally shared similar ethos, open call workshops brought in diverse perspectives and rich conversations. Workshops that were part of existing modules or retreats saw motivated participants and reduced the burden of engagement.

However, designing these distributed sites of engagement was not an easy task. My efforts were limited by existing awareness and possible entry points into this engagement. Only people who had some insight into the effects of predictive tools were motivated to participate. As such, I was unable to reach those who were farthest away from an understanding of AI. I was also aware of the burden of engagement the workshops put on the participants, especially open call workshops that could be organized far away from publics' place of residence or work.

6.1.3 Other inspirations

There is a scarcity of sites that allow for scalable or systemic approaches to achieve transparency [88]. Some scholars studying data literacy such as McCosker et al. suggest that we must intervene in 'organisational data settings', sites where data is developed and used,

especially by community organizations and non-profits who work directly with communities. This, they urge, will help address any 'expertise lag' that may occur when working with data [218]. Pallett et al. propose the development of *observatories* to mediate public engagements with civic AI systems. The use of observatories will allow XAI researchers to take advantage of existing civic systems designed for public engagement. They can also be placed in a manner that allow ongoing conversations in response to the continuously growing worlds of AI. These observatories can also be designed to deliberately focus on under-represented communities and mediate conversations with diverse and distributed communities [219].

Current approaches that mediate community engagement with civic AI tend to do so by partnering with non-profit civic organizations [8, 115, 220]. Other approaches involve creating new spaces of interactions via open workshops [221], engaging online via thorough documentation of AI systems and their effects [131], or developing frameworks that communities can utilize to regulate technologies [222]. All these methods offer ways to think about how we can create sites for explaining with predictive tools that not just welcome but prioritize the voices of those most harmed by predictive tools.

6.2 Designing explaining media

What forms can explaining take? When designing systems for explaining, XAI researchers must consider the media through which explaining and knowing can happen. Effective explaining media can support publics in accessing and understanding AI systems in relation to their lives and contexts.

Most public-facing explanations of AI systems exist as white papers documenting their workings, websites documenting the products, media articles describing the products, their use, and their effects, or in forms of proprietary information in institutions that may (or may not) be accessible via (Freedom of Information Act) FOIA requests. Most of these forms are difficult for publics to access and include languages that require people be skilled in

understanding these systems with respect to their goals [4]. Some forms that may be easy to access may be too simplistic and may rarely, if ever, provide useful explanations. Much information about predictive systems may be undocumented and unpublished, and may not even exist in forms that can be retrieved by the public.

6.2.1 Our approach, learning, and challenges

In the beginning of this research, I, as a member of the public, attempted to engage with existing modalities that offer AI explanations including locating relevant papers and reports, filing FOIA requests, researching media articles and contacting journalists, and analyzing the works of other scholars and activists who have studied place based predictive policing. I hit several roadblocks. Rarely, if ever, the technology makers documented their processes for public access. My FOIA requests were denied and I had little legal knowledge to challenge authorities. When reviewing public articles, I encountered contrasting information where some information may be dated or incorrect. For example, while one article reported the suspension of Predpol in Atlanta, another said it was still in use. A journalist I spoke with as part of the study in Chapter 2 empathized and declared that she had to stop studying and writing about these tools due to a lack of access to understandable information. I was unable to access people and institutions who may know more about these products. The few people I did hear back from, including a now-retired police officer in East Atlanta, knew little about the workings or use of these tools. These efforts to access existing media through which explanations of AI systems could be consumed lasted for around 4 months and produced limited results. The little information I gained by reviewing academic [91] and activist work [131] were then translated into languages of my designed media. I believed that my design of media would promote access and understanding for diverse publics. This work employed two primary media to situate the explaining and knowing in publics' contexts: mapping and workshops.

Mapping. I used maps/mapping as tools/techniques to engage with the workings of

predictive systems. Maps offered a tool to think about AI systems in grounded ways that were centered in the neighborhoods where participants lived and worked. Participants were familiar with the history and culture of places on the maps in ways that helped them imagine how predictive systems may categorize spaces. Additionally, grounding the discussion in participants' lives allowed them to identify spatial, systemic, and social factors that will influence the workings of the predictive systems. The discussions were documented on a large shared map using markers.

Participatory Workshops. Explanations were created and shared through participatory workshops. In my preliminary research involving interviews with communities, participants suggested the use of websites, videos, online lectures and panels, as potential media. These media, people suggested would allow the transfer of information to be easier to distribute, more efficient to consume, and easier to access. However, even though these forms already exist, publics remain unaware of the workings of civic AI systems. These forms are also not interactive. Lastly, they put us, as researchers, in the position of being 'experts' and having the ability to explain these systems to communities based on their needs.

To address these limits, my work employed participatory mapping workshops guided by a loose protocol to co-develop systemic explanations of AI. I designed for shared power in the development of explanations and actively considered everyone in the room an expert who brings in their local knowledges to help us collectively understand the workings and impacts of predictive systems.

6.2.2 Opportunities and challenges

Using maps as the medium guiding the explanations allowed us to understand "AI in place". Maps helped spatially organize the thoughts of diverse audiences. They kept the explanations grounded in spaces that participants knew and lived in. As we would expect to see in participatory workshops, participants led the conversation often building on each other's

thoughts or challenging them. Participants brought in diverse local knowledges to pluralistically explain AI [223]. Bringing those knowledges together, on a shared map, resulted in the group being able to compare different neighborhoods in relation to the disparate algorithmic effects they may feel.

Both these media helped develop grounded critical thinking in participants and workshop organizers. However, they had their limits. Sometimes participants were not strongly connected to their neighborhoods and therefore were not able to draw on their knowledge of the spaces they lived in. They may be unaware or new to a space. Very often people preferred talking about spaces with maps as a reference instead of marking their explanations on the map. Marking on the map seemed time-intensive and would break the flow of conversation. This made it difficult to document their discussions and partial explanations. In the participatory workshops, at times I was placed in an authoritative role where the participants looked at us to educate them. I tried to reset this power imbalance by informing participants of our goal to collectively understand predictive systems. Other times, participants came in with strong opinions and ideas about the role of AI in society. It was not always possible to engage with those ideas productively, especially if it meant challenging them. The explanations were limited to the fixed time, space, and capacity of the workshops [224] and the constraints of the maps and its projected data layers.

6.2.3 Other inspirations

XAI research need not merely focus on providing neat, structured explanations of complex entangled predictive systems [136]. The form and medium of explaining can reflect and provide opportunities to engage with the complex ways in which predictive tools organize cities [21, 10]. Understanding AI systems may happen through a wide variety of media including the use of simulations, models, or tinkering with a tool (where the tool itself is the medium). Even though these techniques may not always result in accurate explanations of how a given AI system works, they can represent AI systems in relatable ways making

them easy to understand [225]. Visualization, as well as physicalization of AI, have been proposed as methods that can support critical engagement with AI systems [226].

More recently, XAI researchers are calling for explanations to move away from being individual-centric to being community-centric. Some researchers have shown how explanations can be 'socially constructed' such as through conversations on community forums that support users in collectively explaining how an AI platform works and affects them [189] . Even in the adjacent field of data literacy, researchers have called for, project-based learning and hands- on peer-learning [227]. An increasing number of researchers are calling for interactive explanations or dialogical explanations [54, 55]. Such explanations allow users to ask follow-up questions to an automated agent.

Both techniques of mapping and participatory workshops were motivated by my goals to ground explanations in publics' environment (via mapping) and city's broader environment (via participatory workshops).

6.3 Designing explaining interactions

What interactions support practices of explaining? The interaction techniques and protocols that guide the process of explaining and knowing affect the creation of situated and systemic explanations. Interfaces, and the interactions they support for explaining, must consider who gets to explain, the constraints they encounter, and their effects on our collective understandings of AI systems.

Public facing explanation of civic AI systems tend to appear as a one-way transfer of information through webpages or news articles. The dis-embodied nature of these interactions treats the goals of explanations as accessing information, not knowing or understanding AI systems. It also does not support the broader move towards promoting the public understanding of science and AI by producing a 'literate citizen' [228]. These interactions aim to prescribe the feelings of trust or fear instead of promoting independent critical thinking or training people to seek out relevant explanations.

6.3.1 Our approach

Personating. The workshops invited participants to personate the working of the AI system and mark on a map the spatial predictions they think an AI system may make, i.e. participants were invited to 'step into the shoes of place-based predictive policing' and categorize spaces as 'high crime' or 'low crime' on a map. In the pilot sessions, some participants reported that they felt 'uncomfortable' making predictions as they realized the socio-political assumptions they made and the discriminatory stereotypes that drove those assumptions, for example assuming that low wealth neighborhoods may see high crime, thereby labeling poor people 'criminals'. This helped them understand the emotional responsibility a decision-maker feels in making these decisions and if and how AI systems skirt this essential responsibility [177].

Later, the workshop groups discussed why they categorized places a certain way in the personation exercise. As they engaged in this exercise, they reasoned about how an AI system may work by reflecting on their predictions. For example, when assigned the task of predicting, many participants first asked 'what kinds of crime are we talking about', which led to the group discussing how AI system's predictions may differ based on the kinds of crime they focus on. They were asked to imagine how their reasoning for marking places as high crime may be reflected in existing data sets and how these datasets may or may not serve as *true* reflections of ground reality. Such discussions led to people discussing the pros and cons of 911 data, arrest data, crowdsourced data, and socio-economic data. They were asked to then consider how their predictions divide space and the physical boundaries they create. Participants noticed that some of their predictions were at the scale of an intersection, while others were at the scale of neighborhoods. They considered if and how AI tools divide spaces in ways that may cause irreparable damage, not unlike the racial segregation created by red lining maps in in the 1930s [229]. Participants also considered the time-scale of data that went into their own predictions. They realized that for some areas, their predictions were based on much recent data, while for other areas, which they

perceived as 'hardly changed', their data was from a decade ago. Sometimes, they realized that their predictions were based on older data that may not be valid because of the growth and development in the neighborhood. Lastly, many participants talked about how these tools affect spaces in real life. They reflected on the effects of labeling spaces as 'high crime' on the economic investment that an area attracts, the people who live there or frequent the area, and the relationships of people with each other.

Protocols. Even as I attempted to abide by the ethos of participatory research, I realized the need to have prompts that can help guide the conversation. These prompts guided the personation process described above. I noticed in the pilots, that the more components of the AI system I revealed, the better the participants were able to relate it back to their expertise. It supported participants in identifying their expertise domain in relation to the predictive tool. The protocol involved having people map out the places they thought would be marked as high crime by an AI system (as described in the personation section above). At times when the conversation slowed down, I introduced components such as prediction type (What is the tool predicting?), Data Type (What kinds of data is the tool using?), Data Selection (How much data and from what sources is being used?), Data Aggregation (How is the data being aggregated in space?), and Prediction Impacts (How does the prediction impact space). This is detailed in the toolkit shared in the Appendix C. I may not have the accurate explanation of an AI component. For example, I may not know what kinds of data are fed into place-based predictive policing systems or what the source of the data is. Yet, merely introducing this as an aspect that informs AI workings allowed the group to speculate, understand, and explain how these tools may work in their neighborhoods and how they mat affect their communities.

6.3.2 Opportunities and challenges

Personating AI helped participants understand AI workings by reflecting on the decision-making and reasoning that implicitly happens when they themselves make predictions.

Such reflection helped identify questions and concerns AI systems may have to engage with as they make predictions. Personating required workshop coordinators to provide prompts that triggered the reflective process. These prompts consisted of revealing AI components that inform AI workings. These were loosely followed but were essential to guide the participants to think about their prediction in relation to AI. Personating AI predictions focused on the socio-technical workings of AI—what happens that allows AI to make predictions. We engaged in additional speculative exercises to consider how AI systems impact space—what happens after AI makes predictions.

Our prototcols were flexible and participants at many instances introduced algorithmic components that may be more relevant or concerning to them. The protocol allowed us to introduce explanation points (tiny pieces of information similar to data points) that the participants could then relate to in their local contexts. There were struggles. The protocols developed from one workshop to another as I learned what AI components are most relatable. However, since this work was not done in close collaboration with one social group, I may not have offered appropriate prompts to diverse groups and their local contexts.

6.3.3 Other inspirations

Some works introduce interactions that follow the lead of Seymour Papert who proposed 'learning by doing' [230]. Participatory Art has been proposed as a strategy to support public interaction with civic, often 'invisible' AI [231]. Other modes of creative interactions such as through the use of activity boxes [232] have shown promise in explaining AI.

Visual languages including familiar signs, symbols, and icons have been used to make AI legible in accessible and user-friendly manners [233]. Additionally, thought experiments such as forward engineering AI systems, in contrast to reverse engineering, have been proposed to proactively help us understand algorithmic impacts [234]. Critique as a mechanism of 'AI transparency' has helped uncover discriminatory power structures that form the foundation of AI [235].

Other projects such as the Moral Machine by MIT [236] provide users with a platform to imagine how AI system could or should work by taking the place of AI and making judgments about the movement of an automated car in 'trolley car'-like problems. Personation has been considered an effective method for understanding the workings of other complex technological systems by others. Seymour Papert developed the coding language LOGO to teach students maths and computation by following the perspective and trail of a turtle on a screen. Students were able to draw on their knowledge of their own bodies to imagine how a turtle would move [237]. Another example is discussed in an ethnography written by Janet Vertesi who demonstrates how scientists used gestures and the materials around them on earth to make sense of how the Mars Rover would see its environment and operate on Mars. They used this sense to code the rover's movements [238].

We believe that a thorough consideration of the design dimensions noted above—designing sites, designing media, and designing interactions— can support the creation of systems of explaining that are situated, systemic, continuous and partial. The reader of this dissertation may have noticed that I have not yet discussed the fourth quality of good enough explanations that I conceptualized in Chapter 2, i.e. good enough explanations are actionable. So, how can good enough explanations be placed in sites, shared through media, and consumed through interactions, that promote public action? This has been a difficult question to answer in the course of this PhD with limited time and resources. Yet, I share some of my reflections below and discuss paths forward.

6.4 Good enough explaining goals and following public actions

In Chapter 2, I found that good enough explanations are not good enough by themselves. They are good enough in so far as they can support public action. *How then can explaining support public action?*

Through this research, I realized that publics do not always have specific goals or actions that motivate their search for explanations. Participants in the workshops noted in

their survey responses their desire to learn more about 'AI' or make sense of their concerns as their *goal* (detailed in chapter 3). Oftentimes, a lack of knowledge and understanding may prevent the conceptualization of effective public actions. Publics may not be aware that a situation necessitates an action if they are not aware that there is, in fact, a 'situation' made up of civic AI tools that is affecting them in invisible ways. Introducing partial explanations, however, including simply making publics aware of the presence of place-based predictive tools, initiated publics' desire to act. I observed instances of this in my work. A participant in Workshop 1 who works as part of a police diversion group, stated that now, having learned how predictions work, they want to aim their team's diversion efforts to places predicted as having 'high crime' with the forethought that the people there may be most in need of their group's assistance. Another participant in Workshop 3 wanted to share learnings from the workshop in civic meetings to guide civic funding decisions. Some educators in Workshop 5 discussed the need to build critical AI literacy for their students in this age of rapid AI growth. As such, one of the goals of good enough explaining can be to support the formulation of publics' desire to act, transforming passive citizens into active and literate systems required for the functioning of a healthy democracy [239].

Another observation was the publics' desire to come together in a room to talk about predictive systems, share concerns, and learn. Dazzled by the growth in AI around them, they welcomed sites where they can gather and talk in an organized but low stakes manner. This was felt by the organizers not only during the workshops that happened and the discussions that preceded and followed them, but also through the workshops that did not happen. I was in talks with four other groups in Atlanta who recognized the need to come together to learn about AI and were enthusiastic about participating. One organization proposed long term collaboration to offer such workshops to the community members they work with as part of the 'academic' pillar and another offered to collaborate to design panel-like learning experiences for policy-makers. Unfortunately, despite my excitement and desire, I was unable to organize additional workshops due to resource constraints. What I did see however

was the formation of publics around these issues. That, I believe, is another communal action that explanation sites can achieve. As Le Dantec explains in his book 'Designing Publics' [240], publics can take shape over time in the course of collective design practices (in our case collective explanation practices). Even though the timeline of this work was much shorter, I see *designing publics* around responsible use of civic AI systems as an action that good enough explaining can attempt to achieve.

In the course of the workshops, I also heard participants express their inability to *act*. A participant in Workshop 1 asked if "all we can do is vent". My position as an individual researcher separate from the civic system (choice of site) gave me limited ability to provide the participants with direct pathways to action. Louder *venting*, however, can be seen as a necessary start.

Another project that I contributed to during my PhD offers a valuable example of a site designed to promote public action. The project designed a curriculum called 'Youth Advocacy for Resilience to Disasters(YARDs)' that used Mapspot [174] to conduct participatory mapping with young BIPOC students. The week-long program, conducted as part of summer camps, supported middle school students in drawing on spatial data to understand the effects of natural disasters on their communities. Next, it prompted students to propose infrastructural improvement projects to build resilience in face of natural disasters. Lastly, it provided them an opportunity to advocate for changes by presenting their proposed improvements to local experts and public officials. The program deliberately designed a site to promote *action* such as advocating for infrastructural change. Such interactions with local and city leaders can be built into explaining sites to deliberately promote public action.

The dissertation work is limited in its ability to promote direct public action towards redesigning, challenging, or discontinuing AI systems. I was also unable to study long term effects of the workshops on public action. The scope of the work was limited to promoting public actions such as reflection, discussion, and debate. In future work, I would like to encourage deliberate formulation of public actions and document ways in which

participants can directly act in service of their goals through civic, systemic, or public channels.

6.5 Guiding future research: Questions to consider

This dissertation calls for the need to *bring partial explanations together, slowly but continuously, to diverse audiences especially those most at risk, in service of their diverse goals*. To that end, below I provide a list of questions that the XAI community can ask as they design for explaining civic AI systems to diverse publics.

Situated:

- Who is aware of and has access to an explaining **site**? How can awareness and accessibility be promoted or channeled towards those most at risk of AI harms?
- What affordances does an explaining **medium** offer? Whose situated knowledges and learning needs are privileged or discounted and how?
- How do explaining **interactions** mediate the process of knowing to meet diverse, situated, and long-term critical thinking needs? How do they support publics in drawing on their own expertise to effectively question or explain predictive systems?

Systemic:

- How does the explaining **site** allow for the space, time, and resources needed for systemic exploration of predictive systems? Who funds and supports these sites?
- What tools and languages does the explaining **medium** provide diverse publics to aid them in identifying, communicating, comparing, and organizing their partial explanations?

- How do the **interactions** allow diverse publics to identify components of AI systems they interact with, are knowledgeable of, or seek to learn about to promote systemic explaining and knowing?

Slow and partial:

- How does the explaining **site** support long-lasting interactions for continued explaining and knowing?
- How do explaining **media** allow gathering, organizing, and continuing slow explaining and knowing?
- How does the explaining **interaction** they support allow partial, but thoughtful creation and consumption of explanations? Does it account for changes in public perceptions and knowledges with time?

Actionable:

- What connections and networks does the **site** support that are in service of the defined explaining **goals**?
- How do the explaining **media** along with the **interactions** allow publics to know and explain in ways that helps them formulate, communicate, and initiate necessary local actions for themselves and their communities?

CHAPTER 7
CONCLUSION

> "Without awareness about these kinds of technologies, and without some level of understanding about how they're being used, there's no possibility for democratic oversight...Political change comes from getting individual people, regular citizens, involved in their local communities, and without an understanding of the existence of these tools, of who is making decisions about how to procure them and use them, how they're operating and so on, people can't mobilize around injustices that the tools help to perpetuate." Participant from Chapter 2

Today, even as AI systems rapidly invade every aspect of public life, everyday citizens remain unaware of what predictive tools exist, how they work, and what their impacts are. Without such an understanding citizens remain ill-equipped to inform, challenge, or protest the use of AI tools even as they are constantly subjected to their effects. They may not know why their job application was rejected, why they were not given a loan, why their child did not get into a school, or why they are seeing an seeing an increase in police presence in their neighborhood. This leaves citizens dis-empowered to demand just and equitable outcomes from civic decision-making organizations. Ultimately, it hinders citizen's democratic right to understand the systems that affect them, vocalize concerns, exercise meaningful participation, and hold civic organizations accountable.

How then can we support public understanding of the workings and effects of civic predictive systems? In this dissertation, I conduct qualitative and participatory research to suggest one possible approach to address this question: designing systems for 'Good enough explaining'. Good enough explanations, as I theorise in Chapter 2, may not be complete or universal, but are good enough to support diverse publics in critically engaging with the design, use, and regulation of AI. Such explanations (1) are *situated* in the

lives of diverse publics, (2) explain the complex and entangled socio-technical *systems* that predictive tools interact with, (3) involve *continuous and partial* processes, and lastly (4) empower publics to *act* in ways that promote democratic deployment and regulation of predictive tools.

Accepting the proposition of good enough explanations, one may ask: how do we then instill these qualities in explanations? To study this, I employ a 'research through design' approach and examine how good enough explanations may come to be (see chapter 3).

'Situated' explanation needs, as I demonstrate in Chapter 4, emerge from how publics relate to algorithmic contexts through (1) the predictive domain: the service domain that an AI model becomes a part of— in our case policing, (2) the prediction subject: the people or places that are subject to predictions—in our case neighborhoods, (3) the predictive backdrop: the local and global environment that surrounds predictive systems, and (4) the predictive tool: the tools and models that make predictions. These relations, I argue, must be considered in our attempt to design explanations for the evolving and situated needs of diverse publics.

'Systemic' explanations of the socio-technical assemblages surrounding AI, that may address 'situated' explanation needs, need not be developed by the XAI community isolation. Instead, as I demonstrate in Chapter 5, these can be co-created with local publics who are positioned well to partially explain how algorithms interact with society to affect local contexts. Locals directly engage with components of AI systems outside of the 'black box' and can make known, amongst other aspects, the environments AI tools are deployed in, the cultures and norms they invade, and the lived experiences of the problems they attempt to address. XAI researchers can support the design of spaces that bring local publics together to help uncover nuanced networks underlying complex AI systems.

'Continuous and Partial' processes can support the development of 'situated' and 'systemic' explanations. AI systems can never be known in their entirety. As with all knowledge, they can only ever be known from specific standpoints and in specific settings [195],

i.e. they can only be known and explained *partially*. Continuously evolving AI systems, therefore, demand continuous, long-term processes for the development of partial explanations. To moderate such processes, as I demonstrate in Chapter 6, the XAI community will have to grapple with three primary design decisions: (1) designing explaining sites (2) designing explaining modalities, and (3) designing explaining interactions. These decisions will allow for the development of situated, systemic, continuous, and partial explanations.

Lastly, 'Actionable' explanations are surrounded by mechanisms that allow social groups to act. Unfortunately, I, as an independent researcher in academia was unable to study long term effects of the workshops on public action or promote direct public action towards redesigning, challenging, or discontinuing AI systems. However, my work highlights the ability of *good enough explanations* to support publics in formulating actions, i.e. considering what public actions are necessary for the responsible deployment of civic AI tools and what could be their role in service of those actions; and formulating collectives that can come together to sustain processes of learning, debate, and action.

7.1 Context Considerations and Limits

While I have identified considerations, constraints, and limits of this research in the previous chapters, here I provide three overarching points.

First, this research is limited in its scope. I ground my investigation in the use of one public safety predictive tool— place based predictive policing. Researchers have studied this tool for over a decade which makes it a strong case study for investigating the development of effective explanations of this tool. Additionally, it allowed me to investigate geospatial workings of AI which are underexplored by the XAI community. Yet, many emerging and 'hyped' tools such as AI Chatbots (ChatGPT) and Facial Recognition tools regularly came up in conversations. A more generalized conceptualization of public understanding of AI requires more work.

Second, this research focuses on the use of civic AI in the United States (US) and the

workshops are conducted in Atlanta, GA. As a developed nation in the Global North, my learnings from a study conducted in the US will not translate globally. More work is needed to understand the unique explanation needs of publics in developing countries that are also increasingly developing and deploying AI tools for public safety amongst other service domains.

Third, the recruitment of interview and workshop participants was motivated by my goal to learn more about public explanations of civic AI from a wide range of stakeholders. I aimed to recruit participants with diverse experiences and relationships with predictive policing systems across their areas of participation and expertise. In this dissertation, I do not report on demographic information about the participants, because this dissertation serves as exploratory research that focused on social roles emerging around predictive technologies, rather than personal experiences. In later work, I plan to explore the importance of gender, race, age, and other demographic distinctions in the creation of explanations for civic AI.

Ultimately, this dissertation serves as a starting point to investigate public explanations of civic AI systems and I propose concepts that can be useful for my fellow researchers to build on, nuance, or challenge.

7.2 Future Work

Moving forward, this work can be developed in the following ways: (1) iterating workshops for the needs of specific social groups and developing a guidebook, (2) studying long term effects and needs, (3) developing a medium that organizes partial explanations over longer periods, and (4) taking steps to promote public action.

7.2.1 Workshops and guidebook

In the future, I plan to work more closely with fewer local organizations to understand their explanation needs and goals in contextual and grounded ways. As I have mentioned ear-

lier, revealing algorithmic components can prompt the development of explanation needs and help formulate new collective understandings. A close collaboration will allow me to introduce novel and constantly evolving algorithmic components, navigate developments, and identify long-term effects and related actions.

Through more nuanced collaborations and my work in this dissertation, I plan to develop a guidebook (such as the Atlanta Community Engagement Playbook [241]) for community organizers to aid them in conducting these workshops with the communities they work with.

7.2.2 Longer term engagement

Another dimension along which I want this work to progress is longer-term engagement with communities. I want to conduct follow-up workshops with social groups. Through this process, I also want to understand if and how these workshops are affecting their everyday lives.

Gradually over these longer term enagagements, I would also be interested in inviting diverse groups together, such as through open call workshops, as they are more empowered and confident in their expertise to discuss, debate, and demand collective public action.

7.2.3 Organizing partial explanations

In this work, I used 'paper' to map our discussions. While the use of such a medium reduced the barrier to entry and invited more participation, it was also temporary. Often maps tore in transportation and their size made it difficult to look at multiple of them concurrently.

Following the lead of projects such as the Anti Eviction Mapping project [242], I plan to develop a digital, accessible, and widely shareable medium that can allow partial explanations to exist together on the same platform. It will also allow us to visualize these explanations alongside institutional data layers such as census data about race or income

to understand how explanations for spaces differ with such socio-political characteristics. Such a representation was also requested by some of the workshop participants.

7.2.4 Promoting Action

Lastly, I plan to offer workshops in ways that support publics to *act*. I may do so by either offering the workshops in collaboration with existing civic systems that provide access to policymakers or city leaders or by directly including pathways to action as part of independent workshops.

I also see an opportunity to center communities interaction with AI policy [243]. We can do so by engaging in collective workshops with policy makers and community leaders to support each group in explaining these predictive tools in relation to themselves to the other group and formulating collective actions. This would be a longer-term undertaking and would require building robust relationships with multiple social groups.

7.3 Positionality Statement

I identify as a woman born and raised in India, a developing country in South East Asia. I have a background in Human-Computer Interaction. Growing up in a country where safety for women and girls has been a rising concern, I have always been attentive to technological advancements for safety. However, through my education and research, I have also been exposed to the harms these tools can cause even as they aim to promote public good. As a non-tech expert myself, I am attempting to understand what publics (like my own self) would need to know about these tools to carefully and responsibly identify the effects of these systems in relation to my communities and other marginalized peoples.

During the time of the research, I have lived and worked in the United States as part of the Georgia Institute of Technology, a well-reputed university in South USA. My affiliation with the university gives me credibility and may have affected recruitment of stakeholders and participants for my dissertation. I have lived in the city in which the study

was conducted for 6+ years which has helped me better relate to the local contexts and neighborhoods some participants in my research talk about. Ultimately, my analysis of the interviews and workshops I conduct is influenced by my personal and professional background and motivations as described above.

AFTERWARD

Following the completion of my research, I have reflected on my choice to focus on place-based predictive policing as a case study. I believe my choice of this case affected my research in fundamental ways. Here I document what motivated my selection of the case, how the case supported my work, the challenges I faced, and my learnings. My hope is that these reflections can help other students and researchers think about how to approach such selection of cases as they study the design of public centered explanations or technologies more broadly.

Place-based predictive policing was unique as a case study for studying effective explanations due to several reasons:

Availability and access to information. Place-based predictive policing systems have been in development and use for over a decade now. They have been thoroughly studied by activists and academics who have attempted to understand the workings and effects of such systems. The breadth of work done on such systems led me to believe that it may be logistically possible to access information about these systems and therefore explain these systems to diverse publics.

However, I realized in my research that despite the long-standing existence and awareness about the harms of these tools, how these tools function or become part of civic processes still remains unknown. These black-boxed systems are being protected by black boxed institutions such as by the technologists making these tools and policing institutions using these tools. Stakeholders in our study suggested that such protection is intentional. They explained that the more information that is available about these tools for public oversight, the more these organizations are exposed to backlash related to legal and human rights violations. The opaqueness of these systems was also made apparent by the hundreds of Freedom of Information Act (FOIA) requests filed by journalists that were declined. I,

myself, attempted to fill out requests for information about the use of these tools in Atlanta. I used existing FOIA requests as samples and requested information about the AI tools used by the Atlanta Police Department and related white papers, audits, data, or handbooks. My request was quickly declined. This was surprising to me as similar requests, such as the ones I used as samples, had been fulfilled in other states or contexts. In fact, a small part of the information I requested was already shared as part of another public FOIA request. The decision to comply to my specific request was ultimately at the behest of the responding officer. And with little legal knowledge about how to navigate the FOIA process or challenge the decision, I was unable to do much about their choice to decline my request.

Such opaqueness has also discouraged further investigation of these tools. A journalist I spoke with during my work explained how she has now stopped writing about predictive policing systems as they are too difficult to access. Another stakeholder, a police officer that I reached out to, mentioned the use of predictive policing tools in Atlanta but was unable to share anything else about how it worked or the protocols surrounding the tool. The officer also did not respond to repeated requests for further conversations. This was not surprising as other police officers and agencies I had reached out to in the past had never commented on the tool. I am not sure if they themselves did not know about these tools or were unwilling to talk about these tools due to systemic fear, lack of incentive/time, or loyalty towards their departments and associated practices. I believe investigations of such systemic barriers to transparency are highly important to help make civic AI systems known to publics.

Ultimately, while it was difficult to access these systems, my investigation of this specific case helped me understand how civic systems work (or do not work) to limit public access to civic AI tools.

However in my particular case, due to limited access to information about these systems, I employed speculative personation as an activity to collectively develop good enough explanations. One could argue that the explanations about AI tools that emerged out of this

activity are in fact *speculations* themselves and may not be grounded in hard facts. Even as I, the moderator, attempted to ground speculations in algorithmic components through the prompts in the workshop protocol, there was no effective way to identify whether the speculations were in fact the 'reality'. Yet, the process proved useful in three ways: (1) it helped participants develop a critical understanding of how they *expect* AI systems to work even if they did not know how they *do* work; (2) it allowed participants to *explain* the socio-technical assemblages surrounding AI tools in good enough ways. These explanations, unlike the speculations that they may have emerged from, were grounded in local experiences and knowledges, and (3) it helped identify, not what is already known, but what we need to know about the technical or socio-political elements of an AI system to eventually develop a good enough understanding. Ultimately, speculative personation acted merely as a means, a method of doing and imagining, to eventually develop good enough understandings.

Investigating spatial effects. Place-based predictive policing systems provided a unique opportunity to investigate the *spatial* effects of civic AI tools, that remain understudied. As I discuss in my Introduction, several other civic AI tools affect spaces, through spatial variables like zip codes or proxy variables like income or race. However, place-based tools center their very predictions in 'spaces', attempting to dehumanize spaces as inanimate objects that affect crime patterns in generalizable ways. Place-based predictive systems, therefore, provided a distinctly spatial premise to investigate the spatial workings and effects of AI. Some other place-based tools like disaster relief management tools or market analysis tools may serve as similar cases to study in future work.

Relevance for the site of study. Place-based predictive *policing* systems served as a relevant case for my site of study: Atlanta, GA. Atlanta has a history of segregating communities using urban infrastructure such as highways and railways. More recently, an Atlanta based policing project, called the 'cop city' by its critics, is being criticized for promoting oppressive policing tactics. Atlanta is also known as the 'most surveilled' city

in the United States. Additionally, many of Atlanta's technology-based policing projects are funded by the Atlanta Police Foundation, a private organization that does **not** need to comply to public requests for information or public oversight and assessment. As such, I expected the case of place-based predictive policing would hold importance for the citizens of Atlanta and would encourage them to question, understand, and explain these AI tools. My work received attention and praise from many non-profits and social groups who related well to the case, realized the importance, and volunteered their time to participate enthusiastically. While the workshops were grounded in this particular case, participants often brought other cases to the discussion table including other policing tech (facial recognition or license plate readers) and other everyday use tools (ChatGPT, ads, finger print readers). The workshops allowed the participants to deviate to cases most relevant to them and consequent workshops intentionally engaged with tools that participants were interested in.

In summary, while my selection of this case presented challenges, such as accessing information about these tools, it still allowed me to engage with and study the civic systems protecting the tools. The lack of information on such tools combined with the high impact they have on citizens, further motivates the need for continued investigation of these tools. My case also aligned with my research interests and site and (1) allowed me to study *spatial* AI and (2) served as a case that was important and relevant for my participants in Atlanta. When selecting a future case, I expect to face similar challenges in other civic domains, sometimes more than others, depending on how transparent a domain is and the socio-political climate of trust or distrust that surrounds it. In my future work, I also hope to diversify cases and focus on tools that citizens may directly engage with. Drawing on my findings that demonstrate publics' ability to explain AI tools, I would be interested in how explanations may emerge out of publics' conscious and direct engagement with an AI system.

I hope the XAI community draws on my findings from this work and nuances them by studying diverse cases in relation to publics' diverse settings.

Appendices

APPENDIX A

RECRUITMENT MATERIAL

Here, I share the flyer and webpage that were developed and shared for recruitment on social groups and participants in the workshops.

MAPPING SMART CITIES

This participatory workshop aims to advance our collective understanding of Artificial Intelligence (AI) systems guiding civic processes in our cities. Participants will engage with a mapping toolkit to collaboratively explore the workings and impacts of technologies designed to predict crime in smart cities. The workshop will provide participants with a vocabulary to ask critical questions about civic predictive technologies and promote democratic oversight over these systems.

Thinking through how predictive technologies impact our cities

90 minute workshop, free of cost | Date and time will be determined based on participant availability | No technical background required

Register at https://t.ly/NRVXr or scan:

Georgia Tech

*Questions? Contact **Shubhangi Gupta** at shubhangi@gatech.edu*

Mapping Civic AI Workshop

Thinking through the workings and spatial impacts of civic AI systems

Registration

To register, please fill out this survey or contact Shubhangi Gupta at shubhangi@gatech.edu

Dates

Fri, Nov 10th, 2023, 4:00 PM- 5:30 PM

To request a workshop on another date, please contact Shubhangi Gupta at shubhangi@gatech.edu

Location

In person. Georgia Tech Campus (unless otherwise requested). Exact location will be shared closer to the date of the workshop.

Organizers

Shubhangi Gupta, PhD Candidate, Georgia Institute of Technology
email: **shubhangi@gatech.edu**

Yanni Loukissas, Associate Professor, Georgia Institute of Technology
email: **yanni.loukissas@lmc.gatech.edu**

Acknowledgements

This workshop uses the Map Room technology originally created in 2017 by the Office of Creative Research in partnership with COCA. This technology was later expanded to build a new, research-oriented Map Room on the Georgia Tech campus. Find more details here.

Branding

To share details about this workshop, please feel free to use this flyer.

Georgia Institute of Technology

About the workshop

Artificial Intelligence (AI) tools are increasingly becoming a dominant yet invisible part of civic decision-making. They are being used in public sector systems including child welfare systems, school allocation, policing, disaster relief management, and many more. While these technologies promise to improve objectivity and efficiency in civic decision-making, they have been shown to have inequitable effects on marginalized social groups. In this workshop, we will use participatory mapping to advance our collective understanding of AI systems used in the public sector, specifically policing. To do so, we will examine technologies that aim to predict crime in cities, how they work, and their impacts, both positive and negative. Ultimately, the workshop aims to provide participants with a vocabulary to ask critical questions about predictive technologies that influence civic processes in their cities.

*Note: This is **not** a one-way teaching workshop where the organizers educate the participants. Instead, in this workshop, organizers and participants think together about the workings and impacts of civic AI systems to imagine and shape better civic futures. The participants are encouraged to bring their expertise developed through engaging with their communities and cities to the mapping table.*

90 minutes long.
Free of cost.
No technical background is required.

Format and Agenda

Format

In this workshop, 6-12 participants will gather around a mapping table to map the workings and impacts of civic AI systems. A mapping table is a table that is covered with paper and has a map projected on it using a short-throw projector. The map has the ability to show various spatial data layers such as locations of crime reports, demographic data, redlining maps, etc. Find more details here.

Agenda

Introduction: We will start by introducing ourselves and posing questions we have about civic AI systems. We will then discuss the public safety needs of the city and if and how predictive policing can help address concerns (or not).

Working of AI systems: We will then think through the perspective of an AI system and imagine how it works to categorize different spaces in a city. We will think through the spatial impacts of the decisions made in the design of system along various relevant components.

Conclusion: We will imagine what a successful predictive tool would look like and what it needed to make it happen.

Upcoming and Past Workshops

November 6th, 2023: We conducted a workshop with the staff of Atlanta Regional Commission (ARC) and Neigborhood Nexus .

October 11th, 2023: We conducted a workshop with the Policing Alternatives & Diversion (PAD) Initiative in Atlanta.

September 18th, 2023: We conducted a workshop with undergraduate students, studying Computational Media, at Georgia Institute of Technology.

Thanks for visiting! For any questions, please contact me at shubhangi@gatech.edu.

APPENDIX B

WORKSHOP SURVEY

Here, I share the survey that was used as a registration form for participation in workshops.

Mapping Civic AI Workshop Survey

Start of Block: Default Question Block

Mapping Civic AI workshop invites communities to engage with a mapping toolkit to collaboratively explore how Artificial Intelligence (AI) systems that aim to predict crime in smart cities work and their potential inequitable effects on the lives of city inhabitants. The aim of the workshop is to support participants in asking critical questions about AI systems guiding civic practices.

Please fill out the short survey below if you will be participating in this 90 minutes workshop. No technical background required.

Questions or requests? Contact Shubhangi Gupta at shubhangi@gatech.edu

Name

_____

Email

_____

Role and Organization (e.g. PhD student, Georgia Tech). Write 'city resident' if you are not participating in affiliation with an organization.

_____

What Atlanta neighborhood(s) are you familiar with (i.e. have spent time in)? Please mention as many as you can. If you are not familiar with any neighborhood, please say 'none' in response to this question.

_____

Why are you interested in participating in the workshop? What do you hope to get out of it? Are there any instances in your life that encouraged you to participate in the workshop?

_____

_____

_____

_____

_____

In your own words, how familiar are you with Artificial Intelligence (AI) systems, their workings, and their impacts? How did you develop the level of familiarity?

_____

_____

_____

_____

_____

Please share any existing thoughts or questions you may have about Artificial Intelligence (AI) systems more broadly or about AI used for civic purposes (including public safety) more specifically. What informed your thoughts or questions?

_____

_____

_____

_____

_____

Any other questions or requests? Write N/A if none.

_____

_____

_____

_____

_____

End of Block: Default Question Block

APPENDIX C

WORKSHOP TOOLKIT

Here I share the toolkit that was developed to guide the workshop interactions and was later shared with the workshop participants.

Questions or Comments? shubhangi@gatech.edu

COMPONENTS OF PREDICTIVE TECHNOLOGIES

- What is the tool predicting?
- What types of data are being used for prediction?
- What is the selection of data being used?
- What boundaries in space are being used to aggregate data?
- What are the spatial impacts of the prediction?

IMPACTS OF PREDICTIVE TECHNOLOGIES

How does space shape predictive technologies and how do the predictions impact communities in space?

Toolkit designed by:

Shubhangi Gupta
PhD Student, Georgia Tech

Yanni Loukissas
Associate Professor, Georgia Tech

PREDICTION GOAL

What is the tool trying to predict? How does the prediction impact neighborhoods?
Why do we want to predict crime in space? Does it align with the public safety needs of various neighborhoods? How can predicting crime help neighborhoods? How can it harm neighborhoods? How does the act of labeling a space as a crime hotspot change the space? Is crime a feature of space? Why? What else is a feature of space that needs to be considered?

DATA TYPE

What type of data is the tool using to make a prediction? How do historical and current discriminatory practices impact this type of data?
Does this data show spatial and temporal patterns that can help predict crime in space? Are these patterns a result of discriminatory practices? What other patterns exist in this type of data? What are the systems in place that lead to the collection of this data? How are these systems just or unjust?

DATA SELECTION

What is the selection of data being used by the tool? How does the selection of data impact people and neighborhoods?
What data are included and what are not? Why? What are the data sources? What are the limits of the data sources? What information is difficult to quantify cannot be part of a data entry? How far back in time is the data from? How do spatial changes over time impact the crime predictions? Which spaces are impacted most by data selection processes?

DATA AGGREGATION

What boundaries in space are being used to aggregate data? How does the choice of spatial segregation impact neighborhoods?
Is the crime predicted for a block, a street, a neighborhood? How is the data aggregated spatially? Do the spatial boundaries reinforce existing boundaries of segregation such as redlining, public transportation routes, etc? What new boundaries along crime do the aggregations form and who would these boundaries impact?

PREDICTION IMPACTS

What are the real-world impacts of the tool? How are these impacts disproportionately distributed in space?
What makes the use of this tool successful? How is the success being evaluated? How does the knowledge of crime predictions impact the people who act on the predictions of the tool and their perspective of various spaces? How does it impact the spaces that are being evaluated by the tool? Which spaces suffer when the crime prediction is incorrect and how?

Predictive technologies for public safety

Place-based Predictive Policing

These tools attempt to predict the location and time of future crimes. These predictions are often used to deploy police forces with the hopes of improving efficiency and objectivity in policing.

Use in ATL: A tool named Predpol (now Geolitica) was used by APD from 2013-2016.

Inequitable effect example: The success rate of Predpol in Plainfield, New Jersey was less than half a percent[1]. Predpol predicted 1940 crimes and 11 crimes in two neighborhoods in Plainfield that are less than a mile apart. These neighborhoods had 0% and 63% white residents respectively[2].

Gunshot detectors

These tools attempt to identify the sound of gunshots. The detection is then used to dispatch police officers rapidly and with more accuracy than traditional 911 calls.

Use in ATL: A tool named Shotspotter was used in Atlanta once in 2018 for a year and another time in 2022 for 3 months.

Inequitable effect example: Michael Williams was wrongfully arrested and he spent a year in jail because a Shotspotter analyst changed the classification of a sound from firecracker to gunshot. This classification was the only piece of evidence presented against Michael Williams[3].

Video Surveillance and Facial Recognition

Surveillance cameras are placed around a city to record and detect people that match specific criteria. They are supported by facial recognition tools that can identify individuals by searching a database of 30+ billion images.

Use in ATL: In 2019, Atlanta bought three 1-year licenses of facial recognition tool called Clearview AI. Atlanta has an active initiative called ‘Connect ATL’ through which it has access to over 18,000 surveillance cameras.

Inequitable effect example: Randal Reid was wrongfully identified by a facial recognition system and spent nearly a week in jail because he bore resemblance to a suspect who had been recorded by a surveillance camera[4].

License Plate Readers

These tools consist of computer-controlled camera systems, that read the tags on a car, and checks if there are any previous issues associated with it.

Use in ATL: The Atlanta Police Department scanned 405,815,610 license plates using automated license plate readers in 2019. The number of readers continue to grow.

Inequitable effect example: Brian Hofer’s rental car that was previously stolen and recovered was identified by a license plate reader and led to a guns-drawn confrontation with the police[5].

1. https://www.wired.com/story/plainfield-geolitica-crime-predictions/
2. https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them
3. https://apnews.com/article/artificial-intelligence-algorithm-technology-police-crime-7e3345485aa668c97606d4b54f9b6220
4. https://www.nytimes.com/2023/03/31/technology/facial-recognition-false-arrests.html
5. https://www.nytimes.com/2019/04/23/opinion/when-license-plate-surveillance-goes-horribly-wrong.html

REFERENCES

[1] *Public safety assessment: A risk tool that promotes safety, equity, and justice*.

[2] A. Noriega-Campero, B. Garcia-Bulle, L. F. Cantu, M. A. Bakker, L. Tejerina, and A. Pentland, "Algorithmic targeting of social policies: Fairness, accuracy, and distributed governance," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 241–251.

[3] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 469–481.

[4] S. Robertson, T. Nguyen, and N. Salehi, "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*, 2021, pp. 1–14.

[5] V. Strauss, *Judge calls evaluation of n.y. teacher 'arbitrary' and 'capricious' in case against new u.s. secretary of education*.

[6] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan, "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions," in *Conference on Fairness, Accountability and Transparency*, PMLR, 2018, pp. 134–148.

[7] H.-F. Cheng *et al.*, "Soliciting stakeholders' fairness notions in child maltreatment predictive systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–17.

[8] T.-S. Kuo *et al.*, "Understanding frontline workers' and unhoused individuals' perspectives on ai used in homeless services," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–17.

[9] J. Schoeffer, N. Kuehl, and Y. Machowski, ""there is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1616–1628.

[10] S. Gupta, S. Janicki, P. Casula, and N. Parvin, "Rethinking safe mobility: The case of safetipin in india," in *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development*, 2022, pp. 1–6.

[11] E. P. Goodman, "The challenge of equitable algorithmic change," *Regul. Rev. Depth*, vol. 8, p. 1, 2019.

[12] R. Brauneis and E. P. Goodman, "Algorithmic transparency for the smart city," *Yale JL & Tech.*, vol. 20, p. 103, 2018.

[13] K. Levy, K. E. Chasalow, and S. Riley, "Algorithms and decision-making in the public sector," *Annual Review of Law and Social Science*, vol. 17, pp. 309–334, 2021.

[14] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[15] R. Bartlett, A. Morse, R. Stanton, and N. Wallace, "Consumer-lending discrimination in the fintech era," *Journal of Financial Economics*, vol. 143, no. 1, pp. 30–56, 2022.

[16] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," in *Ethics of data and analytics*, Auerbach Publications, 2022, pp. 296–299.

[17] A. De La Garza, "States' automated systems are trapping citizens in bureaucratic nightmares with their lives on the line," *Time, May*, vol. 28, 2020.

[18] K. Johnson, "Face recognition software led to his arrest. it was dead wrong," *Wired*, 2023.

[19] J. Wiener, *New ice privacy impact assessment shows all the ways the agency fails to protect immigrants' privacy*, 2023.

[20] C. D'ignazio and L. F. Klein, *Data feminism*. MIT press, 2020.

[21] S. Mattern, "A city is not a computer," in *The Routledge Companion to Smart Cities*, Routledge, 2020, pp. 17–28.

[22] T. Schwanenand and M. Kwan, "Critical space–time geographies thinking the spatiotemporal. guest editorial," *Environment and Planning A*, vol. 44, pp. 2043–2048, 2012.

[23] Y. A. Loukissas, "Who wants to live in a filter bubble? from'zillow surfing'to data-driven segregation," *Interactions*, vol. 29, no. 3, pp. 36–41, 2022.

[24] S. Safransky, "Geographies of algorithmic violence: Redlining the smart city," *International Journal of Urban and Regional Research*, vol. 44, no. 2, pp. 200–218, 2020.

[25] S. Zukin, S. Lindeman, and L. Hurson, "The omnivore's neighborhood? online restaurant reviews, race, and gentrification," *Journal of Consumer Culture*, vol. 17, no. 3, pp. 459–479, 2017.

[26] *Lawsuit challenging lapd's refusal to identify echo park cameras*.

[27] N. T. Lee, "Making ai more explainable to protect the public from individual and community harms," 2023.

[28] *The public oversight of surveillance technology (post) act: A resource page*.

[29] W. H. O. of Science and T. Policy, *Blueprint for an ai bill of rights: Making automated systems work for the american people*, 2022.

[30] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[31] H. Schuff, A. Jacovi, H. Adel, Y. Goldberg, and N. T. Vu, "Human interpretation of saliency-based explanation over text," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 611–636.

[32] R. Shang, K. K. Feng, and C. Shah, "Why am i not seeing it? understanding users' needs for counterfactual explanations in everyday recommendations," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1330–1340.

[33] J. W. Vaughan and H. Wallach, "A human-centered agenda for intelligible machine learning," *Machines We Trust: Getting Along with Artificial Intelligence*, 2020.

[34] M. Mitchell *et al.*, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[35] A. I. Anik and A. Bunt, "Data-centric explanations: Explaining training data of machine learning systems to promote transparency," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.

[36] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, 2018.

[37] T. Gebru *et al.*, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[38] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, "The dataset nutrition label," *Data Protection and Privacy*, vol. 12, no. 12, p. 1, 2020.

[39] S. Dhanorkar, C. T. Wolf, K. Qian, A. Xu, L. Popa, and Y. Li, "Who needs to know what, when?: Broadening the explainable ai (xai) design space by looking at explanations across the ai lifecycle," in *Designing Interactive Systems Conference 2021*, 2021, pp. 1591–1602.

[40] M. Díaz *et al.*, "Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 2342–2351.

[41] M. Pushkarna, A. Zaldivar, and O. Kjartansson, "Data cards: Purposeful and transparent dataset documentation for responsible ai," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1776–1826.

[42] J. T. Richards, D. Piorkowski, M. Hind, S. Houde, A. Mojsilovic, and K. R. Varshney, "A human-centered methodology for creating ai factsheets.," *IEEE Data Eng. Bull.*, vol. 44, no. 4, pp. 47–58, 2021.

[43] R. K. Bellamy *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[44] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 56–65, 2019.

[45] S. Bird *et al.*, "Fairlearn: A toolkit for assessing and improving fairness in ai," *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[46] B. Knowles and J. T. Richards, "The sanction of authority: Promoting public trust in ai," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 262–271.

[47] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[48] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[49] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–15.

[50] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: Informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–15.

[51] A. Springer and S. Whittaker, "Progressive disclosure: When, why, and how do users want algorithmic transparency information?" *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 4, pp. 1–32, 2020.

[52] A. DeVos, A. Dhabalia, H. Shen, K. Holstein, and M. Eslami, "Toward user-driven algorithm auditing: Investigating users' strategies for uncovering harmful algorithmic behavior," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–19.

[53] H. Shen, A. DeVos, M. Eslami, and K. Holstein, "Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–29, 2021.

[54] H.-F. Cheng *et al.*, "Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–12.

[55] A. Crisan, M. Drouhard, J. Vig, and N. Rajani, "Interactive model cards: A human-centered approach to model documentation," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 427–439.

[56] C. J. Cai, J. Jongejan, and J. Holbrook, "The effects of example-based explanations in a machine learning interface," in *Proceedings of the 24th international conference on intelligent user interfaces*, 2019, pp. 258–262.

[57] E. Grünewald and F. Pallas, "Tilt: A gdpr-aligned transparency information language and toolkit for practical privacy engineering," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 636–646.

[58] V. Arya *et al.*, "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," *arXiv preprint arXiv:1909.03012*, 2019.

[59] S. Robertson and M. Díaz, "Understanding and being understood: User strategies for identifying and recovering from mistranslations in machine translation-mediated chat," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 2223–2238.

[60] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? ways explanations impact end users' mental models," in

2013 IEEE Symposium on visual languages and human centric computing, IEEE, 2013, pp. 3–10.

[61] L. Hancox-Li and I. E. Kumar, "Epistemic values in feature importance methods: Lessons from feminist epistemology," in *proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 817–826.

[62] U. Ehsan and M. O. Riedl, "Social construction of xai: Do we need one definition to rule them all?" *arXiv preprint arXiv:2211.06499*, 2022.

[63] H. Kaur, E. Adar, E. Gilbert, and C. Lampe, "Sensible ai: Re-imagining interpretability and explainability using sensemaking theory," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 702–714.

[64] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in ai systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–19.

[65] E. Corbett and E. Denton, "Interrogating the t in facct," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1624–1634.

[66] V. Danry, P. Pataranutaporn, Y. Mao, and P. Maes, "Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–13.

[67] Safetipin, *Safetipin, creating safe public spaces for women*, 2023.

[68] L. Kern, *Feminist city: Claiming space in a man-made world*. Verso Books, 2021.

[69] P. Gupta, *India's safetipin app wins asia foundation's lotus leadership award*, Mar. 2019.

[70] Womanity, *Womanity award*, 2023.

[71] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, *How we analyzed the compas recidivism algorithm*, 2016.

[72] A. Sankin and S. Mattu, *Predictive policing software terrible at predicting crimes*, 2023.

[73] M. DeMichele, P. Baumgartner, M. Wenger, K. Barrick, and M. Comfort, "Public safety assessment: Predictive utility and differential prediction by race in kentucky," *Criminology & Public Policy*, vol. 19, no. 2, pp. 409–431, 2020.

[74] B. J. Jefferson, "Predictable policing: Predictive crime mapping and geographies of policing and race," *Annals of the American Association of Geographers*, vol. 108, no. 1, pp. 1–16, 2018.

[75] B. Green, "The false promise of risk assessments: Epistemic reform and the limits of fairness," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 594–606.

[76] V. Marda and S. Narayan, "Data in new delhi's predictive policing system," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 317–324.

[77] C. Rudin, C. Wang, and B. Coker, "The age of secrecy and unfairness in recidivism prediction," *Harvard Data Science Review*, vol. 2, no. 1, p. 1, 2020.

[78] R. Kitchin and T. Lauriault, "Towards critical data studies: Charting and unpacking data assemblages and their work," 2014.

[79] D. Mehrotra and D. Cameron, *The maker of shotspotter is buying the world's most infamous predictive policing tech*, 2023.

[80] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[81] N.-J. Akpinar, M. De-Arteaga, and A. Chouldechova, "The effect of differential victim crime reporting on predictive policing systems," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 838–849.

[82] K. Kirkpatrick, "It's not the algorithm, it's the data," *Communications of the ACM*, vol. 60, no. 2, pp. 21–23, 2017.

[83] D. Demortain and B. Benbouzid, "Evaluating predictive algorithms," *Algorithmic Regulation*, p. 13, 2017.

[84] N. Van Berkel, J. Goncalves, D. Russo, S. Hosio, and M. B. Skov, "Effect of information presentation on fairness perceptions of machine learning predictors," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.

[85] B. Balaram, T. Greenham, and J. Leonard, "Artificial intelligence: Real public engagement," *RSA, London. Retrieved November*, vol. 5, p. 2018, 2018.

[86] F. Eyert and P. Lopez, "Rethinking transparency as a communicative constellation," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 444–454.

[87] T. W. House, *Blueprint for an ai bill of rights: Making automated systems work for the american people*, 2022.

[88] M. Sloane, I. R. Solano-Kamaiko, J. Yuan, A. Dasgupta, and J. Stoyanovich, "Introducing contextual transparency for automated decision systems," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 187–195, 2023.

[89] S. Gupta, "Mapping the smart city: Participatory approaches to xai," in *Companion Publication of the 2023 ACM Designing Interactive Systems Conference*, 2023, pp. 1–6.

[90] D. Harrison McKnight and N. L. Chervany, "Trust and distrust definitions: One bite at a time," in *Trust in cyber-societies: Integrating the human and artificial perspectives*, Springer, 2001, pp. 27–54.

[91] S. Brayne, *Predict and Surveil: Data, discretion, and the future of policing*. Oxford University Press, USA, 2020.

[92] A. G. Ferguson, "Policing predictive policing," *Wash. UL Rev.*, vol. 94, p. 1109, 2016.

[93] D. Purves and R. Jenkins, "A machine learning evaluation framework for place-based algorithmic patrol management," *Available at SSRN*, 2023.

[94] D. W. Winnicott, *Playing and reality*. Psychology Press, 1991.

[95] G. Gerson, "Winnicott, participation and gender," *Feminism & Psychology*, vol. 14, no. 4, pp. 561–581, 2004.

[96] J. Gabrys, H. Pritchard, and B. Barratt, "Just good enough data: Figuring data citizenships through air pollution sensing and data stories," *Big Data & Society*, vol. 3, no. 2, p. 2 053 951 716 679 677, 2016.

[97] Y.-S. Tseng, "Assemblage thinking as a methodology for studying urban ai phenomena," *AI & SOCIETY*, vol. 38, no. 3, pp. 1099–1110, 2023.

[98] U. Bhatt *et al.*, "Explainable machine learning in deployment," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 648–657.

[99] J. W. Vaughan and H. Wallach, "A human-centered agenda for intelligible machine learning," 2021.

[100] Y. Zhou and D. Danks, "Different" intelligibility" for different folks," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 194–199.

[101] J. Metcalf, E. Moss, R. Singh, E. Tafese, and E. A. Watkins, "A relationship and not a thing: A relational approach to algorithmic accountability and assessment documentation," *arXiv preprint arXiv:2203.01455*, 2022.

[102] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," in *Proceedings of the 24th international conference on intelligent user interfaces*, 2019, pp. 275–285.

[103] A. Henriksen, S. Enni, and A. Bechmann, "Situated accountability: Ethical principles, certification standards, and explanation methods in applied ai," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 574–585.

[104] U. Bhatt, M. Andrus, A. Weller, and A. Xiang, "Machine learning explainability for external stakeholders," *arXiv preprint arXiv:2007.05408*, 2020.

[105] Q. V. Liao and K. R. Varshney, "Human-centered explainable ai (xai): From algorithms to user experiences," *arXiv preprint arXiv:2110.10790*, 2021.

[106] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann, "Bringing transparency design into practice," in *23rd international conference on intelligent user interfaces*, 2018, pp. 211–223.

[107] S. Grimmelikhuijsen, "Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making," *Public Administration Review*, vol. 83, no. 2, pp. 241–262, 2023.

[108] R. F. Kizilcec, "How much information? effects of transparency on trust in an algorithmic interface," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 2390–2395.

[109] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. Van Moorsel, "The relationship between trust in ai and trustworthy machine learning technologies," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 272–283.

[110] B. Knowles, J. T. Richards, and F. Kroeger, "The many facets of trust in ai: Formalizing the relation between trust and fairness, accountability, and transparency," *arXiv preprint arXiv:2208.00681*, 2022.

[111] N. Banovic, Z. Yang, A. Ramesh, and A. Liu, "Being trustworthy is not enough: How untrustworthy artificial intelligence (ai) can deceive the end-users and gain their trust," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, pp. 1–17, 2023.

[112] H. Lakkaraju and O. Bastani, "" how do i fool you?" manipulating user trust via misleading black box explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 79–85.

[113] G. Lima, N. Grgić-Hlača, J. K. Jeong, and M. Cha, "The conflict between explainable and accountable decision-making algorithms," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 2103–2113.

[114] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 624–635.

[115] P. Krafft *et al.*, "An action-oriented ai policy toolkit for technology audits by community advocates and activists," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 772–781.

[116] H. Shen, L. Wang, W. H. Deng, C. Brusse, R. Velgersdijk, and H. Zhu, "The model card authoring toolkit: Toward community-centered, deliberation-driven ai design," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 440–451.

[117] M. K. Lee *et al.*, "Webuildai: Participatory framework for algorithmic governance," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–35, 2019.

[118] M. E. Gilman, "Democratizing ai: Principles for meaningful public participation," *Data & Society*, 2023.

[119] H. Warne, L. Dencik, and A. Hintz, "Advancing civic participation in algorithmic decision-making: A guidebook for the public sector," 2021.

[120] A. Colom, *Meaningful public participation and ai*, 2024.

[121] V. Braun, V. Clarke, and P. Weate, "Using thematic analysis in sport and exercise research," *Routledge handbook of qualitative research in sport and exercise*, vol. 1, pp. 191–205, 2016.

[122] H. R. Bernard, *Research methods in anthropology: Qualitative and quantitative approaches*. Rowman & Littlefield, 2017.

[123] J. Slupska, J. Lowrie, L. Irani, and D. Stefan, "How secrecy leads to bad public technology," 2022.

[124] P. Dourish, "What we talk about when we talk about context," *Personal and ubiquitous computing*, vol. 8, pp. 19–30, 2004.

[125] A. Brennen, "What do people really want when they say they want" explainable ai?" we asked 60 stakeholders.," in *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–7.

[126] C. D'Ignazio and L. Klein, "5. unicorns, janitors, ninjas, wizards, and rock stars," *Data feminism*, 2020.

[127] A. Sankin and S. Mattu, *Predictive policing software terrible at predicting crimes*, 2023.

[128] A. Birhane and F. Cummins, "Algorithmic injustices: Towards a relational ethics," *arXiv preprint arXiv:1912.07376*, 2019.

[129] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan, "Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–16.

[130] B. Baykurt, "Algorithmic accountability in us cities: Transparency, impact, and political economy," *Big Data & Society*, vol. 9, no. 2, p. 20 539 517 221 115 426, 2022.

[131] L. Parsons Lab, *Primer on chicago surveillance*, 2024.

[132] P. G. Kelley and A. Woodruff, "Advancing explainability through ai literacy and design resources," *Interactions*, vol. 30, no. 5, pp. 34–38, 2023.

[133] L. Hallnäs and J. Redström, "Slow technology–designing for reflection," *Personal and ubiquitous computing*, vol. 5, pp. 201–212, 2001.

[134] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–21, 2021.

[135] J. S. Park, R. Barber, A. Kirlik, and K. Karahalios, "A slow algorithm improves users' assessments of the algorithm's accuracy," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–15, 2019.

[136] S. Inman and D. Ribes, "" beautiful seams" strategic revelations and concealments," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[137] *Atlas of surveillance: Documenting police tech in our communities with open source research*, 2024.

[138] D. Maass and B. Lipton, *The atlas of surveillance hits major milestones: 2023 in review*, 2023.

[139] S. W. Duxbury and N. Andrabi, "The boys in blue are watching you: The shifting metropolitan landscape and big data police surveillance in the united states," *Social Problems*, spac044, 2022.

[140] L. Thuy Vo, *How we investigated ring's crime alert system for police departments*, 2023.

[141] M. Ananny and K. Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," *new media & society*, vol. 20, no. 3, pp. 973–989, 2018.

[142] L. Adams and S. Burall, *How to stimulate effective public engagement on the ethics of artificial intelligence*, 2019.

[143] N. Selwyn and B. Gallo Cordoba, "Australian public understandings of artificial intelligence," *AI & SOCIETY*, vol. 37, no. 4, pp. 1645–1662, 2022.

[144] R. S. Geiger, U. Tandon, A. Gakhokidze, L. Song, and L. Irani, "Rethinking artificial intelligence: Algorithmic bias and ethical issues— making algorithms public: Reimagining auditing from matters of fact to matters of concern," *International Journal of Communication*, vol. 18, p. 22, 2023.

[145] N. Sheard and A. Schwartz, *Community control of police spy tech*, 2021.

[146] S. Degroff and A. Fox Cahn, *New ccops on the beat*, 2021.

[147] K. Hayes, *Predictive police tech isn't making communities safer — it's disempowering them*, 2024.

[148] J. Snow, "On the mode of communication of cholera," *Edinburgh medical journal*, vol. 1, no. 7, p. 668, 1856.

[149] S. Johnson, *The ghost map: The story of London's most terrifying epidemic–and how it changed science, cities, and the modern world*. Penguin, 2006.

[150] J. Sleigh, M. Schneider, J. Amann, E. Vayena, *et al.*, "Visualizing an ethics framework: A method to create interactive knowledge visualizations from health policy documents," *Journal of medical Internet research*, vol. 22, no. 1, e16249, 2020.

[151] S. Foundation, *Share lab – research & data investigation lab*, 2023.

[152] K. Crawford and V. Joler, *Anatomy of an ai system*, 2023.

[153] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 5686–5697.

[154] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker, "Gamut: A design probe to understand how data scientists understand machine learning models," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–13.

[155] D. Munechika *et al.*, "Visual auditor: Interactive visualization for detection and summarization of model biases," in *2022 IEEE Visualization and Visual Analytics (VIS)*, IEEE, 2022, pp. 45–49.

[156] Z. Wang, J. Ritchie, J. Zhou, F. Chevalier, and B. Bach, "Data comics for reporting controlled user studies in human-computer interaction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 967–977, 2020.

[157] J. Schneider and L. Ziyal, *We need to talk, ai – a comic essay on artificial intelligence*, 2023.

[158] S. Gupta and Y. A. Loukissas, "Making smart cities explainable: What xai can learn from the "ghost map"," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–8.

[159] E. R. Tufte, "The visual display of quantitative information," *The Journal for Healthcare Quality (JHQ)*, vol. 7, no. 3, p. 15, 1985.

[160] E. R. Tufte, S. R. McKay, W. Christian, and J. R. Matey, *Visual explanations: Images and quantities, evidence and narrative*, 1998.

[161] L. Vaughan, *Mapping society: The spatial dimensions of social cartography*. UCL Press, 2018.

[162] R. S. Wurman, "Making the city observable.," 1971.

[163] K. Lynch, *The image of the city*. MIT press, 1964.

[164] A. M. Kim, "Critical cartography 2.0: From "participatory mapping" to authored visualizations of power and people," *Landscape and Urban Planning*, vol. 142, pp. 215–225, 2015.

[165] M. W. Pearce and R. P. Louis, "Mapping indigenous depth of place," 2008.

[166] C. for Spatial Research Columbia University, *Million dollar blocks*, 2023.

[167] Y. Youkissas, *Atlanta map room project*, 2018.

[168] T. O. of Creative Research in partnership with COCA., *St.louis map room*, 2018.

[169] P. Bosselmann, *Representation of places: reality and realism in city design*. Univ of California Press, 1998.

[170] J. Vertesi, "Mind the gap: The london underground map and users' representations of urban space," *Social Studies of Science*, vol. 38, no. 1, pp. 7–33, 2008.

[171] J. Gupta, A. Long, C. K. Xu, T. Tang, and S. Shekhar, "Spatial dimensions of algorithmic transparency: A summary," in *17th International Symposium on Spatial and Temporal Databases*, 2021, pp. 116–125.

[172] J. W. Crampton, *Mapping: A critical introduction to cartography and GIS*. John Wiley & Sons, 2011, vol. 11.

[173] H. R. Lee, S. Šabanović, and S. S. Kwak, "Collaborative map making: A reflexive method for understanding matters of concern in design research," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 5678–5689.

[174] *Mapspot*.

[175] *Digital media demo day*.

[176] *Gvu spring research showcase 2023*.

[177] C. Véliz, "Moral zombies: Why algorithms are not moral agents," *AI & society*, vol. 36, no. 2, pp. 487–497, 2021.

[178] I. Nicenboim, E. Giaccardi, and J. Redström, "From explanations to shared understandings of ai," 2022.

[179] A. E. Clarke, *Situational analysis*. Sage publications, 2005.

[180] U. Ehsan *et al.*, "The who in xai: How ai background shapes perceptions of ai explanations," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–32.

[181] C. T. Okolo, "Navigating the limits of ai explainability: Designing for novice technology users in low-resource settings," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 959–961.

[182] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 8, pp. 2674–2693, 2018.

[183] R. Yu and L. Shi, "A user-based taxonomy for deep learning visualization," *Visual Informatics*, vol. 2, no. 3, pp. 147–154, 2018.

[184] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 11, no. 3-4, pp. 1–45, 2021.

[185] Y. Rong, P. Qian, V. Unhelkar, and E. Kasneci, "I-cee: Tailoring explanations of image classification models to user expertise," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 21 545–21 553.

[186] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, "Interpretable to whom? a role-based model for analyzing interpretable machine learning systems," *arXiv preprint arXiv:1806.07552*, 2018.

[187] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, "Stakeholders in explainable ai," *arXiv preprint arXiv:1810.00184*, 2018.

[188] S. R. Hong, J. Hullman, and E. Bertini, "Human factors in model interpretability: Industry practices, challenges, and needs," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW1, pp. 1–26, 2020.

[189] Y. Kou and X. Gui, "Mediating community-ai interaction through situated explanation: The case of ai-led moderation," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–27, 2020.

[190] M. R. Haque, D. Saxena, K. Weathington, J. Chudzik, and S. Guha, "Are we asking the right questions?: Designing for community stakeholders' interactions with ai in policing," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–20.

[191] *How a pro-trump mob stormed the u.s. capitol*, 2021.

[192] N. Seaver, *Knowing algorithms. in digitalsts: A field guide for science & technology studies*, 2019.

[193] R. Kitchin, "Thinking critically about and researching algorithms," in *The Social Power of Algorithms*, Routledge, 2019, pp. 14–29.

[194] K. Crawford, *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

[195] S. Harding, "Rethinking standpoint epistemology: What is "strong objectivity"?" In *Feminist epistemologies*, Routledge, 2013, pp. 49–82.

[196] L. A. Suchman, *Plans and situated actions: The problem of human-machine communication*. Cambridge university press, 1987.

[197] F. Kroeger, B. Slocombe, I. Inuwa-Dutse, B. Kagimu, B. Grawemeyer, and U. Bhatt, "Social explainability of ai: The impact of non-technical explanations on trust," in *IJCAI 2022 Workshop on Explainable Artificial Intelligence (XAI)*, 2022.

[198] J. Cobbe, M. S. A. Lee, and J. Singh, "Reviewable automated decision-making: A framework for accountable algorithmic systems," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 598–609.

[199] R. S. Geiger, U. Tandon, A. Gakhokidze, L. Song, and L. Irani, "Making algorithms public: Reimagining auditing from matters of fact to matters of concern," 2024.

[200] *The algorithmic ecology: An abolitionist tool for organizing against algorithms*, 2020.

[201] K. Crawford and V. Joler, "Anatomy of an ai system," *Anatomy of an AI System*, 2018.

[202] M. Sloane, "Controversies, contradiction, and "participation" in ai," *Big Data & Society*, vol. 11, no. 1, p. 20 539 517 241 235 862, 2024.

[203] E. Corbett, E. Denton, and S. Erete, "Power and public participation in ai," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2023, pp. 1–13.

[204] K. Blair, J.-R. Leblanc, and L. Oehlberg, "Exploring public engagement with the social impact of algorithms," in *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*, 2019, pp. 129–133.

[205] E. Seger, A. Ovadya, D. Siddarth, B. Garfinkel, and A. Dafoe, "Democratising ai: Multiple meanings, goals, and methods," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 715–722.

[206] J. Barnett and N. Diakopoulos, "Crowdsourcing impacts: Exploring the utility of crowds for anticipating societal impacts of algorithmic decision making," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 56–67.

[207] A. Marian, "Algorithmic transparency and accountability through crowdsourcing: A study of the nyc school admission lottery," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 434–443.

[208] *Ban the scan*, 2024.

[209] V. Braun and V. Clarke, "Reflecting on reflexive thematic analysis," *Qualitative research in sport, exercise and health*, vol. 11, no. 4, pp. 589–597, 2019.

[210] N. Howell, B. F. Schulte, A. Twigger Holroyd, R. Fatás Arana, S. Sharma, and G. Eden, "Calling for a plurality of perspectives on design futuring: An un-manifesto," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–10.

[211] *Npu university*, 2024.

[212] J. Gabrys, H. Pritchard, and B. Barratt, "Just good enough data: Figuring data citizenships through air pollution sensing and data stories," *Big Data & Society*, vol. 3, no. 2, p. 2 053 951 716 679 677, 2016.

[213] J. Burrell, "The field site as a network: A strategy for locating ethnographic research," *Field methods*, vol. 21, no. 2, pp. 181–199, 2009.

[214] Y. A. Loukissas and J. M. Ntabathia, "Open data settings: A conceptual framework explored through the map room project," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–24, 2021.

[215] *My ad center*, 2024.

[216] *Citizen's police academy*, 2024.

[217] S. Gupta, *Mapping civic ai workshops*, 2024.

[218] A. McCosker, X. Yao, K. Albury, A. Maddox, J. Farmer, and J. Stoyanovich, "Developing data capability with non-profit organisations using participatory methods," *Big Data & Society*, vol. 9, no. 1, p. 20 539 517 221 099 882, 2022.

[219] H. Pallett, C. Price, J. Chilvers, and S. Burall, "Just public algorithms: Mapping public engagement with the use of algorithms in uk public services," *Big Data & Society*, vol. 11, no. 1, p. 20 539 517 241 235 867, 2024.

[220] D. Saxena and S. Guha, "Conducting participatory design to improve algorithms in public services: Lessons and challenges," in *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, 2020, pp. 383–388.

[221] E. Justice, *Workshops*, 2024.

[222] P. P. NYU, *Responsible use of policing tech: Evaluative framework*, 2024.

[223] L. Klein and C. D'Ignazio, "Data feminism for ai," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 100–112.

[224] D. K. Rosner, S. Kawas, W. Li, N. Tilly, and Y.-C. Sung, "Out of time, out of place: Reflections on design workshops as a research method," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016, pp. 1131–1141.

[225] A. Páez, "The pragmatic turn in explainable artificial intelligence (xai)," *Minds and Machines*, vol. 29, no. 3, pp. 441–459, 2019.

[226] M. Ghajargar *et al.*, "From" explainable ai" to" graspable ai"," in *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*, 2021, pp. 1–4.

[227] C. D'Ignazio and R. Bhargava, "Creative data literacy: A constructionist approach to teaching information visualization," 2018.

[228] W. H. Hu and R. Singh, "Enrolling citizens: A primer on archetypes of democratic engagement with ai," *Data and Society*, 2024.

[229] D. Aaronson, D. Hartley, and B. Mazumder, "The effects of the 1930s holc "redlining" maps," *American Economic Journal: Economic Policy*, vol. 13, no. 4, pp. 355–392, 2021.

[230] S. Papert, "The children's machine: Rethinking school in the age of the computer," *New York*, 1993.

[231] K. Blair, P. Hansen, and L. Oehlberg, "Participatory art for public exploration of algorithmic decision-making," in *Companion Publication of the 2021 ACM Designing Interactive Systems Conference*, 2021, pp. 23–26.

[232] D. Long *et al.*, "Fostering ai literacy with embodiment & creativity: From activity boxes to museum exhibits," in *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, 2023, pp. 727–731.

[233] J. Lindley, H. A. Akmal, F. Pilling, and P. Coulton, "Researching ai legibility through design," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[234] P. Polack, "Beyond algorithmic reformism: Forward engineering the designs of algorithmic systems," *Big Data & Society*, vol. 7, no. 1, p. 2 053 951 720 913 064, 2020.

[235] T. Hollanek, "Ai transparency: A matter of reconciling design with critique," *Ai & Society*, vol. 38, no. 5, pp. 2071–2079, 2023.

[236] E. Awad *et al.*, "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.

[237] S. A. Papert, *Mindstorms: Children, computers, and powerful ideas*. Basic books, 2020.

[238] J. Vertesi, "Seeing like a rover: Visualization, embodiment, and interaction on the mars exploration rover mission," *Social Studies of Science*, vol. 42, no. 3, pp. 393–414, 2012.

[239] P. J. Ballard, A. K. Cohen, and J. Littenberg-Tobias, "Action civics for promoting civic development: Main effects of program participation and differences by project characteristics," *American Journal of Community Psychology*, vol. 58, no. 3-4, pp. 377–390, 2016.

[240] C. A. Le Dantec, *Designing publics*. MIT Press, 2016.

[241] M. Asad, *Atlanta community engagement playbook*, 2017.

[242] *Anti eviction mapping project*, 2024.

[243] Q. Yang *et al.*, "The future of hci-policy collaboration," 2024.