

‘You are you and the app. There’s nobody else.’: Building Worker-Designed Data Institutions within Platform Hegemony

Jake M L Stein*

jake.stein@cs.ox.ac.uk

Department of Computer Science,
University of Oxford
Oxford, UK

Reuben Binns

Reuben.Binns@cs.ox.ac.uk

Department of Computer Science,
University of Oxford
Oxford, UK

Naman Goel

naman.goel@cs.ox.ac.uk

Department of Computer Science,
University of Oxford
Oxford, UK

Vidminas Vizgirda

s1750767@ed.ac.uk

University of Edinburgh
Edinburgh, UK

Jun Zhao

jun.zhao@cs.ox.ac.uk

Department of Computer Science,
University of Oxford
Oxford, UK

George Chalhoub

george.chalhoub@cs.ox.ac.uk

Department of Computer Science,
University of Oxford
Oxford, UK

Nigel Shadbolt

nigel.shadbolt@cs.ox.ac.uk

Department of Computer Science,
University of Oxford
Oxford, UK

Max Van Kleek

max.van.kleek@cs.ox.ac.uk

Department of Computer Science,
University of Oxford
Oxford, UK

Rui Zhao

Rui.Zhao@cs.ox.ac.uk

Department of Computer Science,
University of Oxford
Oxford, UK

Wael S Albayaydh

wael.albayaydh@cs.ox.ac.uk

Department of Computer Science,
University of Oxford
Oxford, UK

ABSTRACT

Information asymmetries create extractive, often harmful relationships between platform workers (e.g., Uber or Deliveroo drivers) and their algorithmic managers. Recent HCI studies have put forward more equitable platform designs but leave open questions about the social and technical infrastructures required to support them without the cooperation of platforms. We conducted a participatory design study in which platform workers deconstructed and re-imagined Uber’s schema for driver data. We analyzed the data structures and social institutions participants proposed, focusing on the stakeholders, roles, and strategies for mitigating conflicting interests of privacy, personal agency, and utility. Using critical theory, we reflected on the capability of participatory design to generate bottom-up collective data infrastructures. Based on the plurality of alternative institutions participants produced and their aptitude to navigate data stewardship decisions, we propose user-configurable tools for lightweight data institution building, as an alternative to

redesigning existing platforms or delegating control to centralized trusts.

CCS CONCEPTS

- Human-centered computing → Collaborative and social computing systems and tools; Empirical studies in interaction design.

KEYWORDS

Critical/Activism/Ethics, Workplaces, Personal Data/Tracking, Participatory Design, Empirical study that tells us about how people use a system

ACM Reference Format:

Jake M L Stein, Vidminas Vizgirda, Max Van Kleek, Reuben Binns, Jun Zhao, Rui Zhao, Naman Goel, George Chalhoub, Wael S Albayaydh, and Nigel Shadbolt. 2023. ‘You are you and the app. There’s nobody else.’: Building Worker-Designed Data Institutions within Platform Hegemony. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI ’23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 26 pages. <https://doi.org/10.1145/3544548.3581114>

1 INTRODUCTION

After a TED talk and a SIGCHI resource, a slickly-produced video featuring Uber designers is YouTube’s third result when searched for “participatory design” (PD). In the video, a user experience researcher breaks down how the method enables Uber to do what HCI methods do best - reveal unknown unknowns, center user

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI ’23, April 23–28, 2023, Hamburg, Germany

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9421-5/23/04...\$15.00

<https://doi.org/10.1145/3544548.3581114>

needs, and identify associated affordances. The needs addressed, however, remain far from the questions like those Ekbja and Nardi asked HCI to foreground just a few years ago [20]. Theirs were questions like: “which social classes are benefited by a technology,” how do technologies help “recover autonomy of production,” and, “what kind of technologies benefit the working poor...migrant workers...and unions?” Certainly, these questions are some of the most relevant to workers who rely on algorithmically-managed platform work systems run by companies like Lyft, Uber, or DoorDash [31, 51, 60, 61]. But why should we expect a corporation to consider such questions, which are not nearly so relevant – and perhaps even antithetical – to their most important stakeholder: their shareholders? Much of the HCI research aimed at making platforms more worker-centered has focused on redesigning existing platforms to be more just, transparent, and fair. We take a different approach. We assert that platforms themselves are not the only force determining the material conditions of platform labor; and thus should not be the only target for HCI. Instead, we assert worker-centered designs should be positioned with respect to workers’ individual working conditions, identities and the social institutions which surround their work. Accordingly, we argue for resisting platforms’ economically oriented incentives not by redesigning fairer versions, or via regulations, but by confronting them with grassroots organization and collective information sharing. We focus on identifying and tackling the HCI challenges faced by these “alternative institutions” to facilitate this action [55].

Existing research has identified the harms that information asymmetries create for worker well-being and the mechanisms by which such asymmetries are cyclically created and maintained via algorithmic management [31, 34, 35, 51, 60, 61]. By centering worker voices using PD, HCI research has also uncovered a broad set of needs, features, and even rights that could ameliorate these harms [35, 47, 62]. Beyond HCI, other scholarship has focused on the obstacles and social challenges faced by workers who seek to foment solidarity in atomized working conditions, engage in collective action, or reclaim autonomy in their work [3, 30, 32, 41, 48]. Current research in HCI has seldom attempted to address this second set of obstacles, usually focusing on interfaces and features, rather than infrastructures and institutions. However, platforms have scarcely any incentive to implement the worker-centered designs HCI proposes, and thus we shift the discourse to focus on how other institutions (e.g. government, unions, or open-source communities) can overcome the enormous costs and challenges of supporting the technical and social infrastructures necessary to maintain worker-centered data institutions.

This study aims to identify and address the social and technical hurdles related to designing and maintaining new worker-centered data collective infrastructures beyond modifying existing platforms. We foreground the voices of workers in this social-technical movement, using participatory methods to investigate what roles will be necessary, what trade-offs between function and privacy should be struck, who should participate in designing grassroots data architectures, and what data subjects could contribute. More specifically, we aim to examine two research questions:

RQ1: To what extent can platform workers independently design future data institutions that counteract information asymmetries using Participatory Design?

RQ2: What roles, stakeholders, capabilities, and implementation challenges play a part in worker-envisioned future data institutions?

We used a participatory design approach to emphasize the importance of directly involving workers in the process of identifying the social and technical infrastructures which best suit their needs. We started by presenting workers with the data schema supplied by Uber which they provide in response to data subject access requests from drivers exercising their rights under the EU General Data Protection Regulation (GDPR).¹ We then led eleven platform workers through three participatory design exercises to 1) elicit their ability to internalize the schema, restructure it according to their needs, 2) reflect on trade-offs between privacy and function, and 3) consider mitigating strategies for conflicting privacy and utility priorities. Finally, once workers were sensitized with an understanding of the work necessary to configure useful data structures, we asked participants to reflect on the stewardship tasks they had completed and to describe the roles, responsibilities, and social institutional structures which workers might need to support the maintenance and continuous adaptation of a worker-led collective data infrastructure. Past research has provided thorough mappings of platform workers’ needs and desired features, but focused on re-designing platform work apps themselves. This work extends this line of inquiry beyond needs and features, asking workers to address the asymmetric power dynamics that surround platform work by creating the novel institutions necessary to challenge them.

We found:

- (1) Workers identified disparate, often mutually exclusive or conflicting visions for collective data institutions, which may pose challenges for adoption.
- (2) The boundaries workers draw for access to data are highly variable and deeply personal, while often being motivated by social and technical assumptions that can be inextricable from their governance decisions.
- (3) Governance relationships are likewise contested and personal for workers, underlining the important role that Participatory Design (PD) plays in centering workers voices but also highlighting technical and institutional challenges like the ‘cold start’ problem or group moderation and ontology building, which would affect any alternative institution.

We proceed by first situating our research at the intersection of literature on platform work, strategies for building counter data architectures, and data institutions like trusts and collectives. Next, we frame our analysis in critical social theory to motivate the pluralist vision of resistance the paper explores. Then, we address methodological considerations in the design of this research. Finally, we discuss our findings and their implications for future data institutions.

¹The schema consists of fields and tables detailing which data is collected and how it is stored. We were supplied with the schema by a drivers’ advocacy organization. It did not contain any personal information of workers.

2 BACKGROUND

In this section, we locate our research at the intersection of three themes within related work: algorithmic management, “data institutions,” and finally critical computing and participatory counter-data architecture building. We depart from algorithmic management literature by emphasizing the heterogeneity of platform workers’ organization efforts, their needs, and motivations for collecting data themselves, rather than attempting to uniformly debias or resist existing algorithmic systems. Further, we echo calls from the HCI community to critique the structural role of capitalism in design methodologies for social computing systems. These calls inform our search for methods which can help workers overcome resource disparities and obstacles in establishing worker-centered infrastructures rather than focusing on enhancing or redesigning platforms [20]. Finally, we point to critical social theory, in particular Laclau and Mouffe’s “agonistic pluralism,” to motivate the use of participatory design exercises to elicit the subjective boundaries, roles, and responsibilities workers might occupy within collective data institutions [33, 37].

2.1 Platform Work, Algorithmic Management and the Complications of Identity

Scholars have outlined the harms and dynamics of algorithmic management’s application to platform work and put forward new designs to right it [6, 23, 48]. Zhang et al. point out algorithmic management’s particularly central role in platform work noting that, “due to the scale and logistic complexity of platform work, algorithmic management has been a necessity since its inception,” also alluding to earlier foundational work on the gig economy [62].

How should HCI intervene in the capitalist power structures that define platform work? Graham categorizes the potential responses into three strategies: “regulate, replicate, or resist” [22]. Within the HCI community, research has primarily focused on a combination of the former two – “replicating,” by offering fairer designs, and “regulating,” by advocating for mandates to implement changes to platforms [28]. For example, Zhang et al.’s application of PD to platform worker rigorously maps workers’ needs to changes in platforms that might radically improve working conditions [62]. However, the practicalities of creating and maintaining systems that resist platforms are rarely addressed, acknowledging the low likelihood of platforms’ cooperation (Turkopticon being perhaps the most prominent exception [26]).

The social sciences on the other hand, have produced dedicated typologies of the power relations around platform work, pointing beyond the features of platforms to social relations, allocations of capital, and the impact of material conditions of labor, social institutions, and personal identity [16, 57–59]. Wood and Lehdonvirta highlight ways that the information asymmetries undercut capacity building and collective resistance [58]. A distributed workplace and faceless, autonomous management system makes cohesion amongst workers hard to muster [1, 23]. Data collection and analysis processes from platforms are opaque, and the potential benefits of grassroots data aggregation are hidden behind the enormous start-up costs of ‘cold start’ problems and initiating network effects.

What’s more, algorithmic management leaves workers fearful of unexplained bans, gamification tactics atomize workers by encouraging competition rather than collaboration, and ruthless data-driven profit optimization leaves workers too financially precarious to risk losing income due to platform retribution [40].

Wood and Lehdonvirta also draw on industrial relations theory to highlight the paradoxical situation of platform workers who find themselves neither in the common position of employment where they may find solidarity in shared working conditions and antagonism against management, nor the autonomy of being fully self-employed, and thus able to take charge of working conditions by being selective with work assignments [58]. They illustrate how these unique pressures give shape to the sort of organization which emerges via “... small acts of participation via the Internet, which entail fewer costs and resources, may remain the most common form of collective action in the remote gig economy” [58, 59].

The contrast of companies’ ability to put off turning a profit while chasing away competition using venture capital funds while workers are squeezed to scantily make a living (let alone organize for better conditions) is not only an asymmetry of information caused by the design of platforms, but one of the broader capitalist systems which enable it. HCI scholarship has emphasized the need to involve worker participation in designing resolutions to these problems. This message is specifically echoed by Wolf et al. and Ekbja and Nardi in their calls to use PD as a means to resist capitalist patterns of exploitation and advance social justice on the scale of social systems [20, 55].

Solving *how* to advocate for workers must also consider the heterogeneity of both the material conditions of work and of workers’ identities. In her ethnographic and legal scholarship, Dubal draws attention to how elements of identity inform worker attitudes towards algorithmic management and how to regain productive autonomy [16, 17]. Dubal points out that while independent contractor status is typically seen as disempowering for platform workers, “immigrant and racial-minority drivers’ approval of their independent contractor status enabled them to exert control over their bodies, to manage their time and transnational lives, and to affirm their sense of dignity as working-class men” [16].

Taken together, the existing research would point to pluralist, flexible solutions that also consider systems outside the bounds of platforms to enhance platform workers’ autonomy, however workers might choose to define it.

2.2 Data Trusts and Other Institutions

Information asymmetries contribute not only to the source of harms to workers’ well-being in algorithmic management, but also undercut collective resistance by atomizing workers, squeezing workers’ income, and in doing so, preventing them from finding common ground or worse, turning workers against one another [39, 48, 61]. In other contexts, data institutions like data trusts, which collect data and steward it under fiduciary trustees, or collaboratives, which create pooled data resources for contributors, or collectives, which aim to counteract powerful actors through voluntary data aggregation of an organizations’ members, provide the potential for counteracting the information asymmetries between computational systems and their stakeholders, while building capacity

among subjects to collect, analyze, and respond to insights from their data [4, 46, 50].

Edwards put forward data trusts as a means to structurally rewrite the terms of who gets to benefit from data [19, 24]. Trusts' promise is first and foremost as a tool to both unlock the potential of publicly held data which ostensibly belongs to the data subjects whom it is about or who help produce it, while also democratizing the benefits that might emerge from the use of data to derive algorithmic systems. The information and power asymmetries between platforms and workers demand a similar democratization. However; the social institutions and material conditions of those who might benefit, create or support such "bottom-up data trusts," bear no resemblance to the established, well-funded institutions early advocates had in mind [13, 24, 41]. Further, while citizens may have strong claims to collective ownership of data within public institutions, the economic relationship between workers and platforms is much more tenuous.

Recent "bottom-up" data collectivization efforts also offer potential solutions, and acknowledge the need for more robust social support. [11, 13, 25, 41]. Delacroix and Lawrence put forward a vision for a plurality of "bottom-up data trusts" [13, 47]. They imagine a variety of smaller institutions that might allow for the independent idiosyncrasies of data to be accounted for, and data trusts to be purpose-built for data subjects' unique needs. Despite their promise, successful examples of bottom-up data institutions are few and far between; and little research has targeted the immense challenges of initiating and maintaining such grassroots, worker-centered data collective infrastructure from both a social and technical point of view. How workers might overcome the institutional and technical hurdles of implementing and supporting such infrastructures remains unanswered.

2.3 Counter Data Architectures

Counter-data architectures aggregate data about a subject already being quantified by a dominating actor, replacing in-built logics of capital or domination with the epistemologies of data subjects. Dombroski documented data collection tactics of low-wage workers, and the role that data aggregation could play within the process of "designing to reconfigure socioeconomic relations" [14]. Calacci and Colclough directly apply the concept to the workplace, enumerating the new ways unions or other workers' organizations could make use of data collaboratives and the roles they might play [8, 10]. Drivers' Seat collaborative in the US and the Workers Info Exchange in the UK have both worked within existing data access laws and their own technical means to enable transparency between platform workers. However; these efforts face difficult odds. In their study of 25 failed social computing projects that attempted to address unfair wages and working conditions, Wolf et al. found three levels of failure: "individual level of technology adoption; two, relational failures (i.e., the anti-labor worker/employer dynamic in the US); and three, institutional or macro-level failures" [55].

Critical computing literature in HCI informs our study in two ways. First, it illustrates the need for focus not only on technological changes within platforms, but also the social relations surrounding platforms and individual workers [38]. Second, it advocates a "dig where you stand" approach – that resistance to capitalist power

structures necessarily takes many simultaneous and conflicting forms and that to be successful workers must "ultimately shape their own histories," via PD [55]. Bottom-up data institutions promise great value to workers in the platform economy, but face a variety of "individual, relational, and institutional," challenges [55]. Together, this motivates our study to confront workers with existing data from the institutional structures in which they work, then to draw on Laclau and Mouffe's critical theory, asking them to imagine what an "agonistic plurality" of smaller-scale social and technical counter-institutions could exist where *they* stand [33].

2.4 Rethorizing Resistance: Hegemony, Agonism, and Autonomy

Laclau and Mouffe's discussion of hegemony refers to the tendency of societies to maintain existing hierarchies of power and to corrupt even resistance efforts against powerful actors [33]. Wolf et al. voice a similar warning with respect to the capitalist influences which creep into human centered design methodologies for social computing efforts [55]. They caution that when "we talk about capitalism as a giant monolith, as the pervasive and only viable economic system, we further reify it – we make it that ever-stronger monolith. This makes it conceptually more difficult to imagine alternative futures" [55]. We argue the same is true of counteracting platforms' power with respect to workers. Designing for changes to platforms which affect all workers who use them universally, or by building highly centralized or generalized data institutions may work against workers' interests. This rings true in Wood and Lehdonvirta and Dubal's empirical work: platform workers do not see themselves as one class or homogeneous voice, and as a result, expecting a uniform consensus around organization may work against their efforts [16, 58]. In Dubal's study, forcing the one-size fits all solution of employee status on all drivers was seen by many as restricting their autonomy, though such a distinction is often seen as a victory. The wide plurality of desires for how collectives might help workers resist was corroborated by the diversity of designs generated by workers in our study [16, 58].

Mouffe and Laclau's model for collective resistance creates space for the necessary remedies of different workers to conflict with one another, while remaining mutually valid. A question remains and provokes our research: *how can this collective form of resistance be applied to social computing infrastructure, where costs inevitably rise when a common consensus infrastructure cannot be agreed upon?*

Indeed, Mouffe and Laclau consider individuals' differences as constitutive of their collective identity, and thus the source of solidarity and autonomy. Mouffe and Laclau's response to the hegemonic tendencies of centralized consensus making is "agonistic pluralism," the notion that to counteract exploitative patterns in capitalism, a plurality of identity-bound instances of resistance are necessary [43]. Scholars of the data economy have seized on this idea as a counterpoint to the centralized control of data, creating "turbulence" to disrupt platforms' central control [37]. Mouffe and Laclau argue that solutions which seek total consensus are the enemy of democracy and individual agency. In this vein, Delacroix and Lawrence already advocate for a plurality of bottom-up data trusts but maintain the model of delegating data stewardship to legally bound representatives to mediate consensus. We would add

that designing new features to be applied across platforms, even with the use of participatory design, might likewise entrench an unwilling consensus among workers, undercutting resistance by privileging majority-rule rather than allowing workers define their own battle lines of resistance and loci of solidarity. In the following methods section, we explain how we designed PD exercises to enable workers to describe their own equally valid, though often mutually exclusive visions of empowerment, from the subgroups they defined to their data access decisions and even in the forms of collective governance they offered.

3 METHODS

Our study used PD exercises to understand workers' vision for collective social and technical infrastructure to support worker empowerment. In the process, we evaluated to what extent workers could participate in designing data structures that effectively balance their many utility, privacy, and economic trade-offs. Our exercises centered around participants' exploration of what data Uber collects about drivers. Reflecting on their own experiences, participants came up with uses for data, and imagined how these uses might be achieved in a collective data sharing setting.

3.1 Developing Participatory Design Exercises

In developing our participatory design exercises, we were inspired by Zhang et al.'s exploration of platform workers' "algorithmic imaginaries," and previous work that foregrounds workers in the design process of tools to reclaim their "productive autonomy" [2, 5, 7, 54–56, 62]. In particular, Dantec and DiSalvo give a critical account of how participatory design can help navigate publics' orientations towards authorities and enact the *infrastructuring of attachments* – the "social and material dependencies and commitments of the people involved" [12]. The original positioning of participatory design as a method to involve workers in the design process to seek changes to their work environments and technologies strongly resonates with our objectives.

We also considered the validity of accessing imaginaries and experiences, particularly given our interest in the influence of real-world privacy and earnings trade-offs workers face [18]. Accordingly, we weighed the benefits of conducting the study as a field experiment, which could quantitatively ground the divergence in preferences of workers situated in the stakes and decision making of workers. We decided against a field experiment given the intrusiveness into work it might require, and recognized that only a more open format could elicit the set of stakeholders, resources and tools workers might use to construct their imagined data institutions. Other methods of participatory design for resistance against algorithmic systems have been put forward [36, 53]. Rather than adopt methods which would aim to audit or debias algorithms, a bottom-up approach addresses the collection, governance and support of data flows with the aim of challenging social relations of work.

After running four internal pilots with other researchers, we narrowed our design down to three exercises and obtained university ethics panel approval for our study. Our study was designed with the goal of ensuring an active role for workers in the design process, and so we obscured examples in our exercises unless workers

requested more detail. We conducted design sessions remotely via Microsoft Teams meetings, using an online white-boarding application to interact with participants, who were given direct access to rearrange graphical representations of data types and fields. We engaged with participants one-on-one to specifically avoid larger sessions or focus groups in which pressures or deliberative consensus making among the workers might obscure their independent preferences, motivations and imagined institutions. Placing emphasis on the individual, subjective imagination of workers was particularly vital for our study given the atomized nature of platform work and the gamification instilled by platforms. Past research corroborates that platform workers prefer to contribute in smaller, personalized settings, specifically anonymous to other workers [58].

Our design exercises simulated data stewardship activities from requirement gathering and schema design to access control. We asked workers to combine fields, make privacy decisions, negotiate competing or conflicting uses of data (for instance wanting to collectively access location data, but not be personally locatable), come up with mitigation strategies like aggregation or obfuscation to resolve conflicting interests, and reason through the governance structures required to put the preferences workers expressed into action. We expected three outputs from the design sessions: an understanding of workers' capacity to perform exercises that mimic the decisions involved in data stewardship, workers' imagined technical and social configurations of counter-institutions that might execute this work, and finally the usefulness of participatory design to generate counter-data infrastructure as responses to the actual data architectures workers aim to counteract.

We began by sensitizing workers to a simplified version of the data set sourced from Uber, asking them to recategorize data types from the Uber schema into sub-categories according to their own preferences. The Uber data contains fifteen categories ranging from device data (such as granular speed or geolocation) to messages with customers and pay information. Uber provides data 26 CSV files in response to driver Data Subject Access Requests (DSARs). Each file contains numerous fields which are grouped according to categories Uber defines. For example "dropoff_lat" in the "driver lifetime trip data," CSV provides the latitude of where a specific drop-off occurred.²

We asked users to assume the role of a data steward with access to an individual workers' data and later, access to a collective of workers' data. Their first task as data steward was requirements gathering, devising, and explaining *metrics* that could be derived from the data types. The goal of this activity is to observe how the participants may make sense of and structure the data as it is provided by Uber, and how they would restructure it themselves.

Then, the participants were invited to conduct an ontology building exercise, combining the necessary data types to calculate their metrics from different CSVs by dragging different data attributes

²A complete example of how Uber structures drivers' data in Subject Access Request Responses can be found at <https://www.workerinfoexchange.org/uber-guidance-documentForallthewayswedisplayedtheschematoparticipants,please see Appendix D, Figure 15.> Our exercises presented this data in the categories Uber provides, and in several simplified formats with the aim of eliciting how drivers would challenge and restructure it to suit their needs.

02 Driver lifetime trip data

This document provides information on the lifetime number of trips. For your convenience, we have provided an additional description in the table below to better help you understand the data that is included in this document.

02 - Driver Lifetime Trips.csv	
Data	Description
uuid	UUID of the trip
driver_uuid	Your Unique User ID
partner_uuid	UUID of the partner of the driver at the time of the trip
vehicle_uuid	UUID of the vehicle
city_id	This field provides the city id that Uber has associated with your account. For example, 18 is London
currency_code	Currency code of Trip Fare payment by rider (e.g., EUR or GBP)
timezone	Timezone of the city where the trip happened
flow	Flow of the driver (e.g., UberX)
source_tag	The origin of the trip if it occurred via an external API's

Figure 1: An example of a data as it appears in a DSAR response guidance document. From Worker Info Exchange®.

(represented by ‘boxes’) into different locations on the virtual whiteboard. We asked workers to pause here to conduct a risk analysis regarding risks that may arise from combining data and to enumerate associated mitigation strategies that could be applied.

The third exercise asked drivers to define access control groups and assign privacy permissions to these groups by dragging data types into their associated categories on the whiteboard. Further we asked participants to apply any mitigation or anonymization strategies and to keep in mind conflicts in privacy permissions based on past combinations of data types they defined.

At the end of the session, we asked participants to reflect on their work as a whole and consider the stakeholders and social infrastructures necessary to support the data stewardship work they had just completed.

Screenshots illustrating these exercises are available in the appendix.

3.2 Recruitment

We recruited participants using a pre-screening survey which we circulated in twelve platform-specific worker Sub-Reddits, Facebook groups, and independent worker forums. We contacted seventeen further worker unions or advocacy groups in the US and UK. Like in Zhang et al., for participants not referred by unions or advocacy groups, we asked participants to send a screenshot of their driver profile to verify their location and participation in platform work. We included workers from both the US and UK as we were curious to test whether factors like local laws, culture, and environment might inform the perspectives and the social institutions that workers imagine. While the fields we exposed workers to came by way of a GDPR request made in the UK, data collection about workers is regulated even less in the US (presently, workers have means to similar access requests only in California). Our exercises sought to understand how workers would configure data institutions should they have access to the data that platforms collect, and we can expect that at least the data listed in the UK would also be collected in the US. This allowed US participants to contribute perspectives shaped by their own working conditions with greater

insight into what data is collected about them. Second, we were conscious that recruitment via online forums, particularly with two-step vetting and the high time-constraints platform workers faces meant a wider recruiting pool would be necessary to achieve a large enough group of participants.

Use of social media forums and a pre-screening survey is common practice among studies concerning the gig economy both to assure a diverse set of participants and avoid false or duplicate form fills [61, 62]. We found several downsides to recruiting in this manner. Many worker social media groups have strict policies with respect to research recruiting, which may conflict with research ethics. On one occasion, an admin requested that page moderators be given preference and greater compensation in participation ahead of group members. Further, we noticed tangible weariness among workers who are bothered by the proliferation of research flyers in community spaces posted both for private sector user acceptance testing and academic research.

We aimed for a diverse and representative sample of participants. Milkman et al. accounts that there is no consensus on the gender distribution of platform workers, pointing nonetheless to anecdotal evidence that women are over-represented in platform work [42]. Gridwise, an app that assists platform workers reports that women only account for 25% of local platform workers [52]. Our pre-screening survey returned 30.8% women (N=20) and 69.2% men (N=45), despite a concerted effort to reach out to advocates for female-led forums and support groups. We reached out to all 65 potential candidates who submitted an interest form, and interviewed 13 participants in total, 11 of which were included in the study analysis. We collected some information about demographics, including platform tenure and primary platforms used for work. Though we did not directly collect information about data literacy, we imputed an indicative level based on workers’ facility with data stewardship tasks and reactions to DSAR data. Supporting quotes for data literacy level of each participant can be found in Appendix B

3.3 Data Capture and Analysis

Online interviews with participants were conducted on Microsoft Teams, and the audio was recorded and automatically transcribed using the built-in Teams recording functionality. We corrected the transcripts manually using the audio recordings where necessary.

After the first three interviews, two researchers independently coded the three interview transcripts with NVivo.12 using open coding to identify concepts, themes, and events. We then negotiated the resulting codes, removing less relevant ones and merging similar codes, to create a shared set of codes. This codebook was used to code the remaining transcripts, making only agreed-upon additions, when new topics were encountered.

After the inductive stage of open and descriptive coding, we constructed top-down thematic codes, with research questions and overarching themes at the top (from “RQ1. To what extent can platform workers independently design future data institutions that counteract information asymmetries?” to “Capabilities” and from “RQ2. What roles, stakeholders, capabilities, and implementation challenges play a part in worker-envisioned future data institutions?” to “Stakeholders” and “Roles”), going down into more

	Country	Gender	Ethnicity	Main platform kind	Platform Experience	Data literacy
P01	US	Male	Asian	Ride hailing	>2 years	Medium
P02	UK	Male	Multiracial	Ride hailing	>2 years	Medium
P03	UK	Male	White	Food delivery	>2 years	High
P04	UK	Male	White	Food delivery	>2 years	High
P05	US	Male	Native American	Ride hailing	6 months - 1 year	Low
P06	UK	Male	Asian	Food delivery	>2 years	Medium
P07	UK	Female	White	Food delivery	1-2 years	Medium
P08	US	Male	White	Food delivery	>2 years	High
P09	US	Female	White	Food delivery	>2 years	Medium
P10	US	Male	Multiracial	Ride hailing	1-2 years	Medium
P11	US	Female	Preferred Not to Answer	Ride hailing	>2 years	High

Table 1: Participant sample demographics, N=11. All information was self-reported, except data literacy, which was determined by the researchers. For supporting information on data literacy, see Appendix B.

granular topics in an intermediate layer of codes (like “Capabilities > Data decisions”, “Stakeholders > Subgroups”, and “Roles > Personal agency”), down to the descriptive codes that fall within the intermediate layer topics (like “Capabilities > Data decisions > Permissions”, “Stakeholders > Subgroups > Location” and “Personal agency > Individual governance > Vote for reps via union”). The full resulting codebook is included in the supplement.

This mixture of inductive and deductive coding allowed us to systematically examine the interview material, while maintaining openness towards new concepts and themes that we were not anticipating. Rubin and Rubin term this a hybrid model, in between a formal pre-determined coding schema and a fully flexible grounded theory approach [49].

In each of the interviews, participants described potential data institution designs. We summarized each design into visualizations using social machine network diagrams inspired by “sociograms” [45], where we document the roles, motivations, stakeholders, and conflict resolution mechanisms that participants imagined for worker-centered institutions. This was done by identifying parts of transcripts coded as “Collective data uses”, “Data access methods”, “Data decisions”, “Subgroups”, “Other stakeholders”, “Individual governance”, “Data sharing & privacy”, “Collective governance”, and “Personal agency”, and diagramming them using Miro and Figma.

3.4 Reliability and Validity

To ensure the robustness of our study, we took measures at several stages. When recruiting participants, we looked for people who were experienced in platform work (at least 6 months) and could prove their worker status, so that they could make meaningful contributions.

The design exercises were planned to test our research questions indirectly, such that instead of relying on participants’ beliefs about what they might choose in a hypothetical scenario, we could ask them about the relevant stakeholders, roles, and decisions with a concrete example context in mind. By introducing scaffolding (reviewing the Uber driver schema categories with a simple categorization exercise) at the beginning of the study, we tried to avoid disadvantaging platform workers with different prior knowledge

of data that platforms collect and ground the discussion in a real-life setting. We also attempted to avoid information overload by simplifying the visual presentation of Uber driver schema data into succinct general categories – we kept extended details about the types of data in each category for sharing if participants asked for clarification.

Four researchers were trained before conducting 1:1 interviews with the participants. In the training, we covered the importance of avoiding leading questions and prompting with examples unless the participants got stuck. If illustrative examples were needed, the same ones were used consistently. All the interviews followed the same semi-structured interview script with predetermined questions and basic flow of exercises, but interviewers had freedom to pursue interesting discussion points.

Over the period of the study, we interviewed 13 participants and discarded two interviews, because we could not verify two participants’ identities as legitimate platform workers. Following these interviews, we revised our verification measures, asking participants to answer additional questions in the pre-screening form about their most and least favourite aspects of the platform they work with. If the answers were still not enough to verify potential participants, we requested them to share a screenshot of their platform worker profile with us. With every change to the study plan, updated ethics approval was sought from the University Ethics Committee before resuming the study.

3.5 Data Protection

The meetings were recorded after obtaining participants’ written consent and after a verbal reminder of their rights (including the right to withdraw from the study without penalty or impacting their compensation). All identifying data has been removed from the transcripts and quotes, participant names were replaced with pseudonyms, and original audio recordings were deleted after concluding the analysis. Participants’ worker profile screenshots were deleted upon verification.

4 RESULTS

In this section we present our findings according to the major tensions among workers’ descriptions of the functions, affordances,

and stakeholder interactions through which new data and social institutions might act. We compare workers' visions for new social data institutions, grouping similar designs into clusters, while emphasizing that their preferences consistently differed depending on aspects of their own identities, priorities and concerns. We observed a broad range of motivations for collectivization and the associated uses for data that workers might expect from data institutions. Workers also defined divergent data sharing preferences for specific types of data, both with respect to the subgroups they might give expanded or restricted permissions to and their modes of governance. Workers also proposed very different mechanisms for collective governance of data, communicating variable trust in existing institutions like unions, government regulators, and platforms. Finally, workers' expectations for how algorithmic systems or platform companies might respond to worker-led data collectivization also had a significant impact on the institutions they imagined.

4.1 Collective Uses for Data

4.1.1 Income optimization. The most common uses for data institutions offered by workers were ways to better optimize workers' own incomes. Participants rattled off ideas from "average wage per hour," "gross income versus what customers were charged," and "average deadhead per hour" (the amount of time spent driving or riding without cargo). P07 said it most concisely, "Wasting time and possible earnings, that's what it's all about isn't it?"

Seeking visibility into one's own income is a straightforward enough expression of individual autonomy, as workers reported metrics might "allow us to plan when to drive and when to take time off." - P10 (US, Driver). Such metrics also might act to alleviate some of the strongest asymmetries exploited by platforms to dictate workers' behaviors, and limit their productive autonomy in line with platforms' interests: "It's frustrating not knowing from day-to-day. Am I gonna make any money today?" -P09 (US, Courier).

It is easy to see how platforms might benefit by keeping workers in the dark about their hourly wages or net profits per mile. Some workers' interest in these metrics (or indeed withholding them from fellow workers) also exposes the latent efficacy of platforms' gamification tactics. These tactics impose hegemonic chilling effects on workers' interest in organization by pitting workers' sense of competition against their collective interests.

"Their systems is predictable. I don't mind that they hide the pay because the formula is so consistent that it's very easy for me to make an advanced decision... if every driver had access to the data, it would make it infinitely harder for me to fully take advantage of that transparency." -P08 (US, Courier).

"The more you share... The more that people are gonna, you know, modify their behaviour accordingly. And if they all start working what you've worked out to be the peak time, then it's not gonna be the peak time for very long" -P03 (UK, Courier).

4.1.2 Data Sharing to Resist Platform Logics. Others were interested in uses of the very same collective data specifically for their potential to subvert or eliminate platforms' gamification tactics and facilitate better working conditions amongst workers:

"I know that a lot of people probably, they don't wanna share that, but if they could be convinced – that information is really meaty, is really important for everybody. New drivers, all drivers, experienced drivers." -P02 (UK, Driver)

Similarly, P06 lamented the safety risks that arise from surge pricing schemes. He noticed that to capitalize on extra earnings in hotspots, "drivers go berserk...get way too rash in their driving." Instead, he offered sharing data would allow all drivers to preempt spikes in demand by collectively coordinating workload distribution, reducing dangerous competition in favor of "equal pay, so equal pay is equal work."

4.1.3 Safety, Well-being and Social Institutions Beyond the Platform. Contrary to personal pay optimization or direct means to subvert dynamic pay changes from the platform, workers also shared uses for data collectives that aim to support workers with safety measures and or fairness of work practices which platform functionalities claim, but platforms in practice fail to fulfill. Indeed, among workers, platforms' access to vast amounts of granular tracking data creates the impression that they should be able to enforce policies which keep them safe and the platform fair, yet they point to areas in which data collectives could meaningfully enhance their working conditions.

These metrics often targeted well-being, safety and mutual aid, while being articulated with particular attention to workers' personal identities. P07, who works mostly for a food delivery platform in the UK expressed that she wanted records of her orders and the time they took from restaurants, so she could validate instances of gender-based discrimination from restaurants; citing that "being a woman my [orders] certainly don't get prioritized." She mentioned that for women, being able to combine data would have pertinence to safety, giving warnings about customers who are "drunk and a bit horrible...or perhaps where the nearest safe spot could be...you could go for assistance for something like first aid" -P07 (UK, Courier).

Similarly, P06 expressed a desire to have detailed records on customer interactions, citing that "it's mostly immigrant people who work in this," he imagines that having greater access to customer complaints would allow workers to rectify issues that came down to failures in communication, rather than poor performance: "And [workers] don't speak very good English at all. So if he were to phone the customer to, you know, get guidance or help on finding the addresses...some customers are really not cooperat[ive]."

Perhaps some of the most direct appeals for solidarity came from workers who imagined data collectives as a means to be connected with one another and form closer ties, recognizing platforms' purposeful atomisation of workers.

"But the nature of this job is that it's isolated and you have minimal interactions with people who you work with, so you'd like to also know who the other drivers are." -P01 (US, Driver).

"So the one of the big issues working for the gig economy with the apps is how blind you are. You are alone completely. You are you and the app. There's nobody else. You don't know the other drivers...and they want

to keep it that way. The more lonely you are, the better. So anything that can cut that, it's good." -P02 (UK, Driver).

4.2 Data Sharing Preferences

Access control is another essential aspect of any data architecture. When reflecting on the categories of data Uber collects, we found that participants had radically divergent sharing preferences, with some imagining data much more freely shared among workers, while others much more protective of worker data.

The general openness to sharing was scattered: P04 chose to share two of fifteen categories of data with Union representatives and no data with other workers. On the contrary, P05 shared all his data either with all workers, or with workers not in his immediate area. The remainder of workers decided to share roughly a third of their fifteen data types with all other workers, withhold a third from everyone, and restrict a third to particular groups³. There was also almost no consistency on the types of data, except for customer-focused data like cancellations and safety reports which all but one participant decided to expose.

When participants chose to share data with specific access restrictions, they proposed a variety of different circles (or sub-groups) to share with.

The main sub-groups which our participants identified were characterized by:

- Income Share – There are people for whom platform work means "minimum wage jobs where it's a starting point for teenagers and young people" and "the people that rely on this for their bread and butter to feed their families ..." -P01 (US, Driver)
- Shared experience – Those that have been working in similar situations, in marginalized groups or at the same time have more in common, for example, "they can bring a way different perspective than perhaps someone at the top that's never been out on a late night delivery and hasn't used any of the apps and doesn't know all the problems you might face" -P07 (UK, Courier) or "Our [group] is really good friends. We became friends because of working in the same store" -P09 (US, Driver).
- Relationship – Human connection not necessarily attributed to shared experiences: "I rely on my Spidey sense more. Uh, so if I had a good rapport with the driver or... uh... You know or or built the relationship with them in some kind of other way. Then I'd be way more willing to give that information [useful data like demand hotspots]." -P01 (US, Courier)
- Shared cultural background – Many form groups where "it's mostly defined by ethnicity" -P04 (UK, Courier).
- Location – workers in the same locations were both more and less willing to share data with those in the same location: "I've heard about horrible stories from London, Birmingham, Glasgow [...] damaging cars and, you know, trying to steal their orders, while they're having a cigarette, or you know, things like that. [...] some drivers are not even welcome in their areas" -P06 (UK, Courier).

³See Appendix C Table 3 for a full breakdown of individual data sharing preferences among participants.

- Transportation method – mostly applying to the context of delivery work, the transport method is also key: "car drivers will talk to car drivers, bikes will talk to bikes. [...] it is highly unlikely that you're gonna get a Portuguese car driver talking to a Pakistani bike rider" -P04 (UK, Courier).

The large number of distinct subgroups poses a significant challenge for design of data institutions, especially because workers had incompatible sharing preferences according to these groups.

With some sub-groupings, workers even had inverted preferences, P10 would only share data with other drivers if they were not co-located, while P07 was only keen on sharing data only with locals. Some even identified the same group: workers with more experience or higher hours worked per week, with P03 determining he would give experienced workers less access, and P01 determining experienced drivers should have greater access.

A further challenge is posed by the most popular kind of relationship workers defined for enhanced sharing - personal relationships. Relying on personal relationships illustrates how data institution designs need to be situated in lives and working conditions of workers and their associated systems orbiting platforms, in addition to the economic or transactional data participants responded to.

The complexity both within individuals' personal preferences, and amongst the group may be a clue into why all-or-nothing approach of trusts to data have been hitherto unsuccessful. The challenge becomes finding the correct social institution which can mediate the common interests and mitigate the group-specific concerns raised by data collectivization. The nuance of workers' specific visions for collective data use and governance also demonstrates that only providing fine-grained controls for data sharing fails to acknowledge the complex institutions needed to balance delivering the benefits of collective cooperation with the preferences of individuals or subgroups. This is another instance in which mistaking information asymmetries as a problem addressable to altering platform functionality is myopic towards the social dynamics which surround and inform platforms and platform work communities.

4.3 Clusters of Proposed Data Collectives

While we focused on imagined collective infrastructures, participants often shared how they thought the existing data sharing ecosystem works. Reviewing all material coded as "experience-based contributions," we summarized them in the sociogram in Figure 2, representing the current landscape of platform work. This serves as a baseline for comparison with the imagined collective infrastructures produced by participants, which have been organized into clusters and can be found in Figure 3-Figure 14.

Participants mentioned a variety of stakeholders both in their description of the status quo and in their imagined institutions. The main stakeholders in the current landscape are platform workers, customers, and the employer platforms that connect them. Governments indirectly play a part via legislation that regulates platform work. Unions are only relevant in some jurisdictions. Third parties that make apps to be used alongside platform work play a sideline role. In the following section, workers described how these stakeholders interact with intermediary data infrastructures.

We clustered proposed data collectives by their main distinguishing features. Some of the solutions could fit into more than one

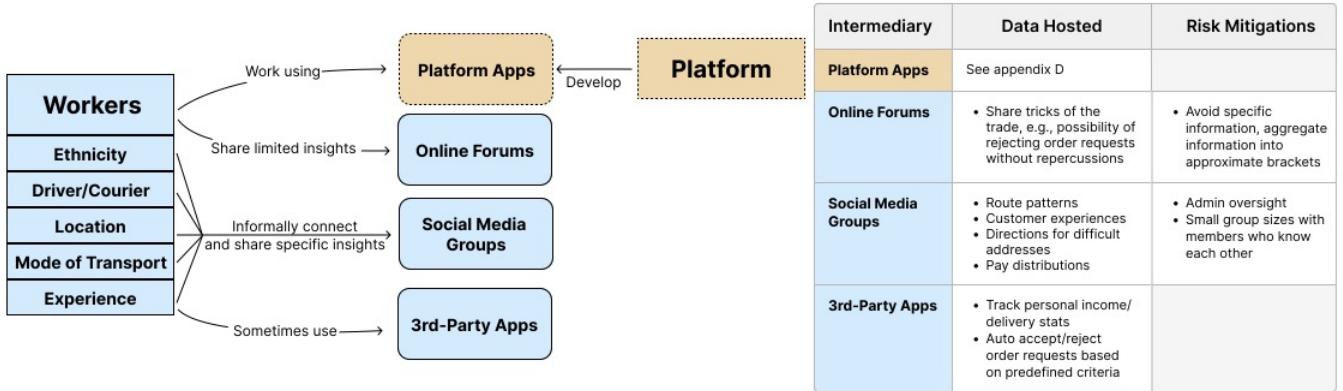


Figure 2: Status quo in the platform work landscape

cluster, in those cases, they are grouped together with most similar diagrams.

4.3.1 “Collective advocacy”. These solutions involve new ways for platforms to gather feedback and workers’ voices to be heard. However, sharing information and making changes is still at the discretion of platforms. Examples of how this could work are shown in figures 3, 4 and 5. Though platforms often have feedback mechanisms, participants’ designs sought insight from patterns in collective rather than individual data to advocate for improvements.

4.3.2 “Blockchain”. P08 proposed a platform-less design, where passengers could “simply use blockchain technology to connect drivers with customers directly, eliminating all the middlemen completely.” The example design is illustrated in Figure 6. A real-life example of this is being developed by Arcade City⁴.

4.3.3 “Collective wiki”. These are designs featuring a Wikipedia-like portal, co-created by platform workers to host data, insights and contextualize information which is openly editable and contestable among workers. Examples are shown in figures 7, 8 and 9. We are not aware of real-life examples of this in practice. Although not required in theory, the examples shared by participants all rely on platforms publishing additional data.

P01 emphasized the need for collective sense making tools in the “Wikipedia-style, where everybody just has access to this spreadsheet or whatever they put their info in,” adding, “If you don’t contextualize it and analyze it doesn’t make any sense.” Another participant stressed the importance of moderation in shared information portals to avoid threats from platforms:

I know Wikipedia somehow magically functions really really well, but, like, I worry about social platforms of any sort get.. the amount of bots that we, like company created bots, that we get flooded with is like really kind of crazy. -P11 (US, Driver)

4.3.4 “Unions on social media”. Solutions in this cluster are characterized by a union (or unions), comprised of volunteer workers, sharing information on social media. The union volunteers bring

in experience from their own platform work and connections, and share useful advice with others for example the participant design depicted in Figure 10. Some unions already do this, for example, the App Drivers and Couriers Union (ADCU) on Twitter.

4.3.5 “New app”. These solutions involve a purpose-built application independent from the platform. The main benefit of these approaches is that they can be used on-the-go by workers, unlike the higher-latency solutions like social media or a collective wiki. Designs involving new apps are illustrated in figures 11 and 12. There are many practical examples of apps developed for various niche uses, for example Flush⁵, which shows public toilet locations; or Surge⁶, which reveals where Uber Surge pricing is active on a map; or Muver⁷, which automatically switches between ridesharing platform apps as rides are completed and new trip requests come in.

One participant shared how an app could show live map data visually, such as high demand hotspots or average earning brackets per area, so that it could be used while driving:

“if you are using one of the apps, you don’t have much time to, kind of, move around too much or to have much more detailed data, so [it] has to be more graphical kind of stuff. And an app is great for that.” -P02 (UK, Driver)

Another participant highlighted that many platform workers are either not native English speakers or not well versed with technology, and for them, solutions with lots of text or complex usage would not work so well. Instead, she stressed that, “you have to assume it’s only being used on mobile [...] It definitely has to be an app. It just has to be, like, ultra simple.” -P11 (US, Driver)

Security was also a noteworthy concern for several participants when considering 3rd party apps, they would only want genuine platform workers to be able to access this data:

“link to your relevant app, like Uber or Deliveroo, so you could sort of sign in or maybe like a private login e-mail that you could go in like a company platform.

⁵<https://play.google.com/store/apps/details?id=toilet.samruston.com.toilet>

⁶<http://www.surgeapp.org/>

⁷<https://play.google.com/store/apps/details?id=taxi.muver.driver>

⁴<https://arcade.city/>

Take some off, get yourself or maybe even like a company web page or something. That would be, you know, it was secure." -P07 (UK, Courier)

4.3.6 "Upgraded platform". One inevitable conclusion for participants was to place governance authority in platforms' hands, given platforms' already entrenched positions. Overcoming the institutional inertia of platforms as workers' primary point of contact for work demands significant investment on the part of collectives. Most importantly, workers like P10 pointed to platforms' legal claim to data, "it should be up to the platform, because [data] belongs to them." P02 explained "whether we want [the platform] to be involved or not is irrelevant." We found the tendency to default to platforms as yet another instance of hegemonic influence – generating institutions also depends on wrestling with workers' existent perceptions of data ownership.

The solutions visualized in figures 13 and 14 propose that platforms should modify their existing apps, adding new features to allow platform workers to see insights from colleagues in the platform app. Interestingly, the participants proposed designs in which platforms are not required to make more data available in the apps themselves – instead, a union or individual workers would voluntarily contribute data to share. This could also be possible with the above "new app" approach, but participants expressed that it was important that data would be available "in the [platform] app itself to make it most usable, if you're talking about people all having access to it..." -P03 (UK, Courier).

4.4 Qualifiers, Risks and Caveats

Many participants shared important caveats for implementations of their proposed data institutions.

4.4.1 Privacy Considerations. Data access permissions decisions might intuitively be conceived as a balance of privacy risk and potential benefit to workers and their peers. Participants generated a wide variety of potential risks, including the disruption of platform dynamics mentioned earlier. P08 cited legal risks of breaking platform terms of service, while P11 brought up risks of retribution from the platform like shadow bans. In general, workers were aware of the risks of sharing data, but were largely unconcerned with their likelihood. P04, a UK courier, noted that platforms already capture intimate details about his everyday activities:

"I'm saying there's a company using an algorithm to track me and send me orders. So I don't really mind that much. If someone was able to track me real time."

P09 expressed similar deference to the risks of sharing data,

"I mean, other than jealousy, like 'oh, her rating is better than mine' [...], I don't think so."

Workers also mentioned techniques such as anonymization, aggregation and perturbation of data shared to make them more secure with collectivization. Anonymization was commonly mentioned as a precondition to data sharing "some things you don't have to share unless it's anonymised like hours you work and your earnings" -P01 (US, Driver). P02, a UK driver, mentioned that "you could fiddle," explaining further that in prior work he would round earnings up or down to obscure exact figures or share within a group pooling

information – akin to perturbation of data in privacy preserving computation techniques.

4.5 Collective Governance & Methods for Accessing Data

Given the divergent visions for data use and management already discussed, finding a structure which can accommodate this plurality of epistemes without forcing a potentially exclusionary consensus is a difficult task. Participants were able to expound on the strengths and weaknesses of governance models and identify stakeholders who might be in positions of control over governance decisions.

4.5.1 Co-determination and Sense-Making, Preferably Without the Trolls. Despite the variability in the size of sub-groupings workers suggested collectives might take, their unique preferences necessitate means for large numbers of workers to contribute to data governance decisions. Workers saw value in the collective moderation and decision-making capabilities of Wikipedia-like open forums. Indeed, the ability for workers to contest and contextualize data was a popular requirement, "The driver should be allowed to comment on it, whether it's private or public...With a public annotations and private annotations." -P01 (US, Driver)

As in other aspects of data institution design, different needs took precedent for different workers. P09 critiqued open forums "because of trolls," then pointed to "voting for representatives" as an alternative. P02 agreed that "we can scratch Reddit off from all of this because they just take the 5% of the most toxic drivers and I think a Wikipedia open platform would be great for information access." On the other hand, there was also plentiful suspicion around delegating decisions to centralized figures like unions or government.

Similarly, some workers suggested personal data stores held separately from platforms and accessible via credentials (not linked to any platforms) as another option. For example, P06 suggested a report with "credible login details or login credentials. Where we could download any quick PDF format." P11 said she "still think[s] an app is exactly the right way to do it, because you have to assume it's only being used on mobile like you know, it doesn't really need to function on a laptop." However, she emphasized Unions' ability to gain the capacity to support such solutions solves both worker participation, and collective governance:

"If a Union sees itself or even if the outside world sees it as an intermediary organization and that is not, I will not recognize that as a labor union. A labor union is of the workers by the workers."

4.5.2 Institutional Slowness and Platform Inevitability. Challenges inherent to data architecture design surfaced with respect to re-using existing institutions like government regulators and unions for collective governance. P06 expressed "I don't want the government to get involved in this. I think they do not understand, because things happen so fast in the gig economy." Others were similarly skeptical of unions' ability to manage technical infrastructure due to resource constraints and requirements to swiftly adapt to changes platforms make. Still some thought unions and regulators would play pivotal roles as moderators or even delegated controllers.

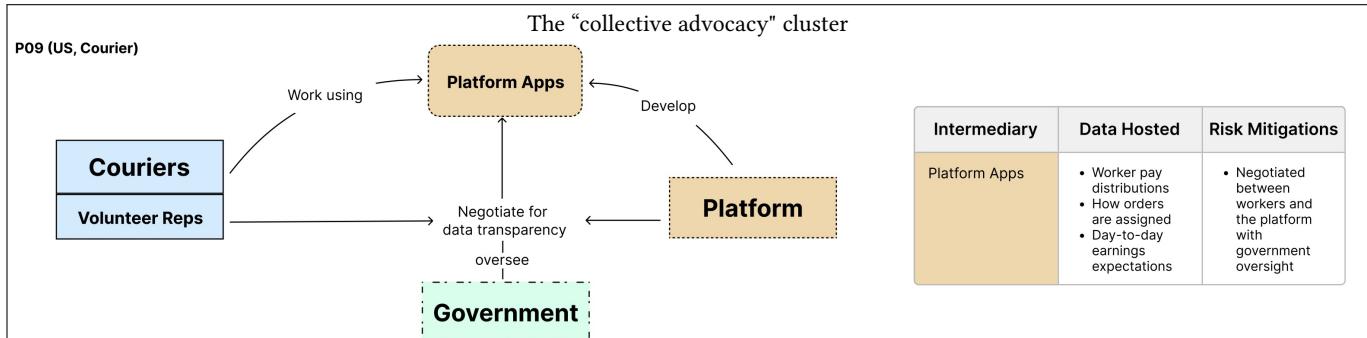


Figure 3: P09 (US, Courier) design

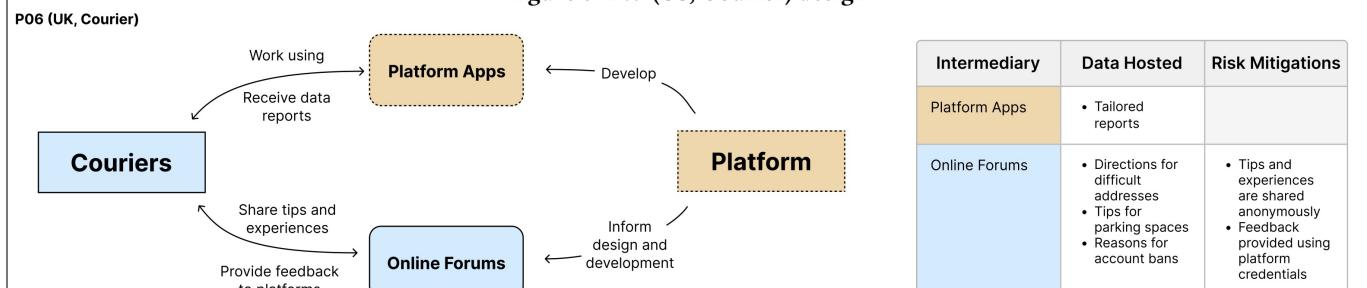


Figure 4: P06 (UK, Courier) design

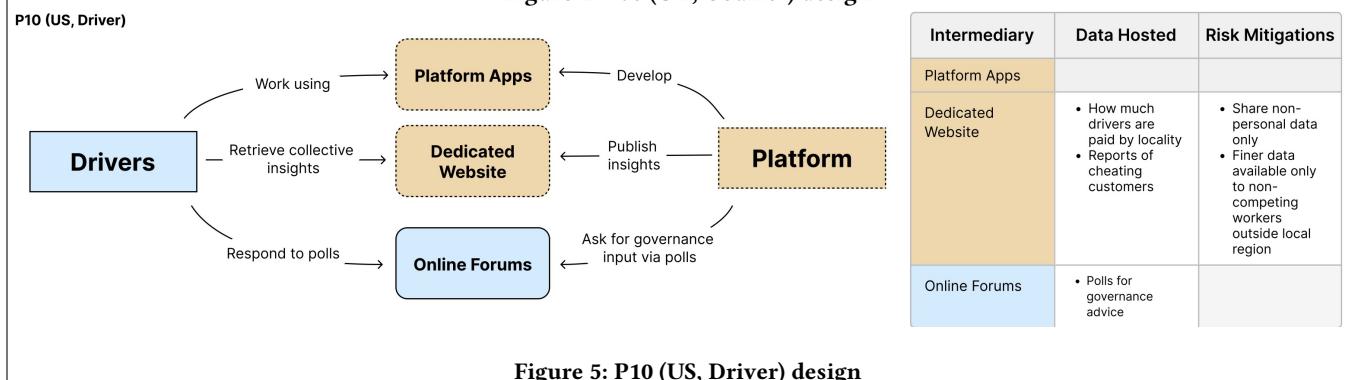


Figure 5: P10 (US, Driver) design

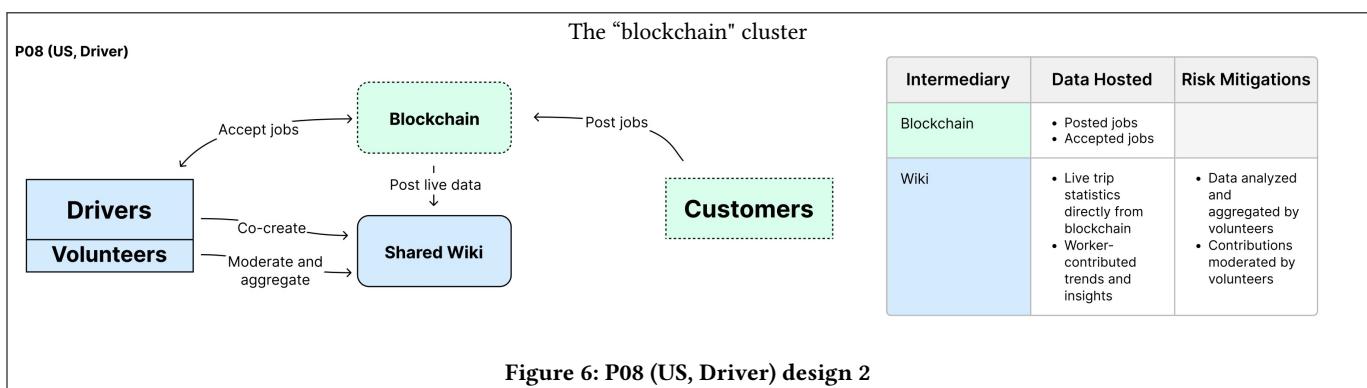


Figure 6: P08 (US, Driver) design 2

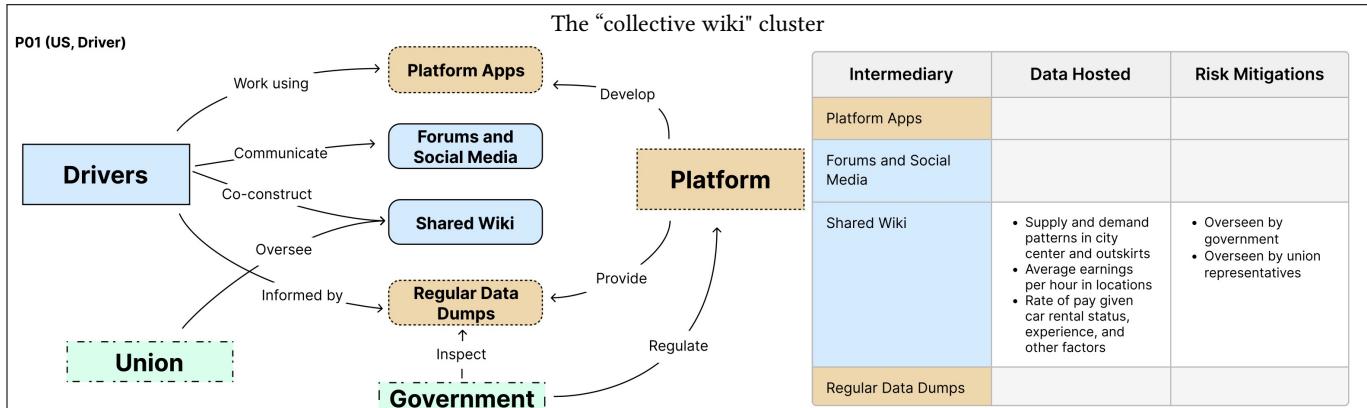


Figure 7: P01 (US, Driver) design

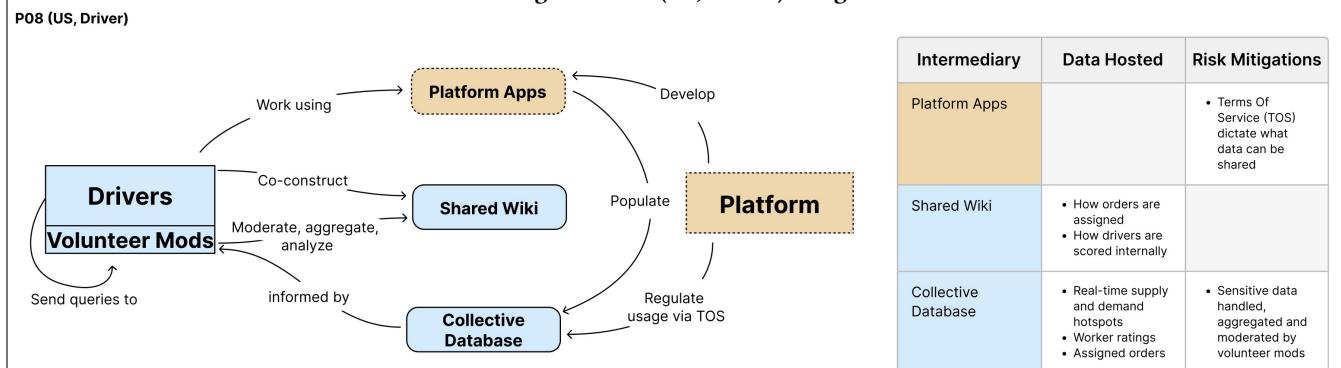


Figure 8: P08 (US, Driver) design 1

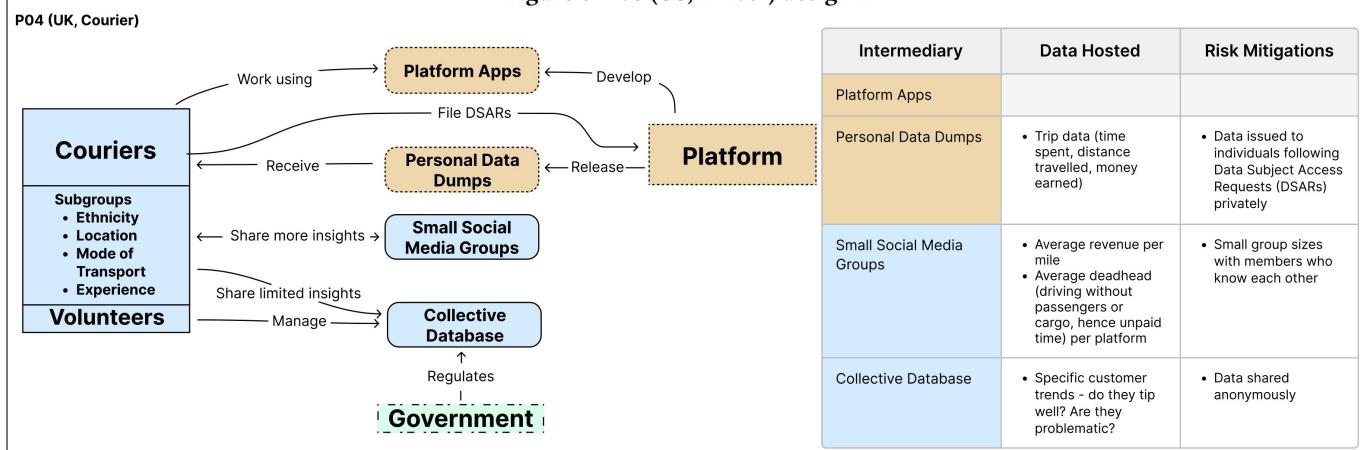
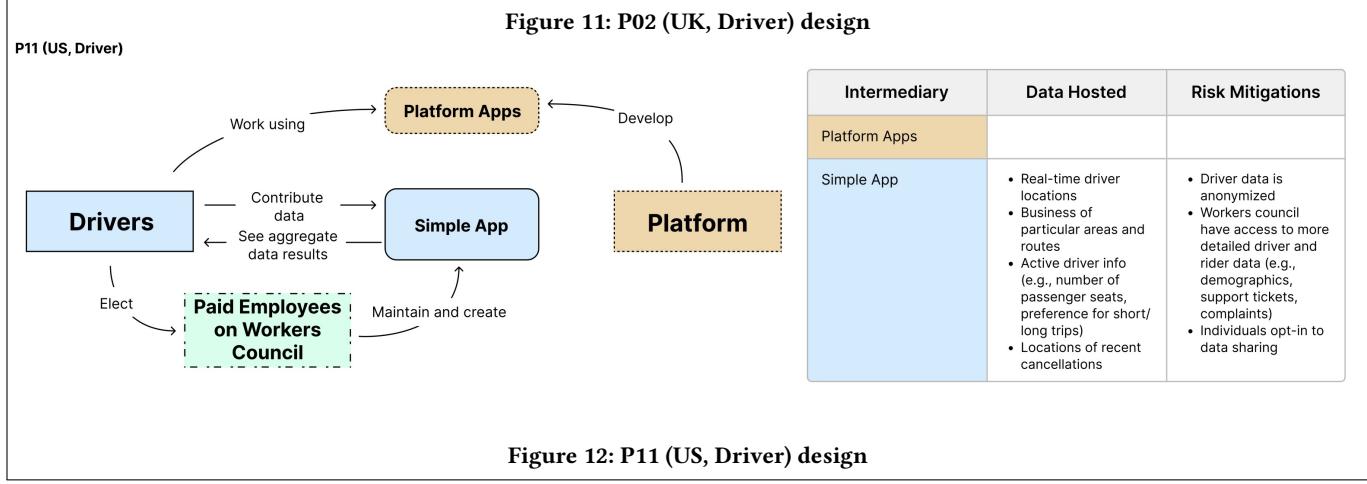
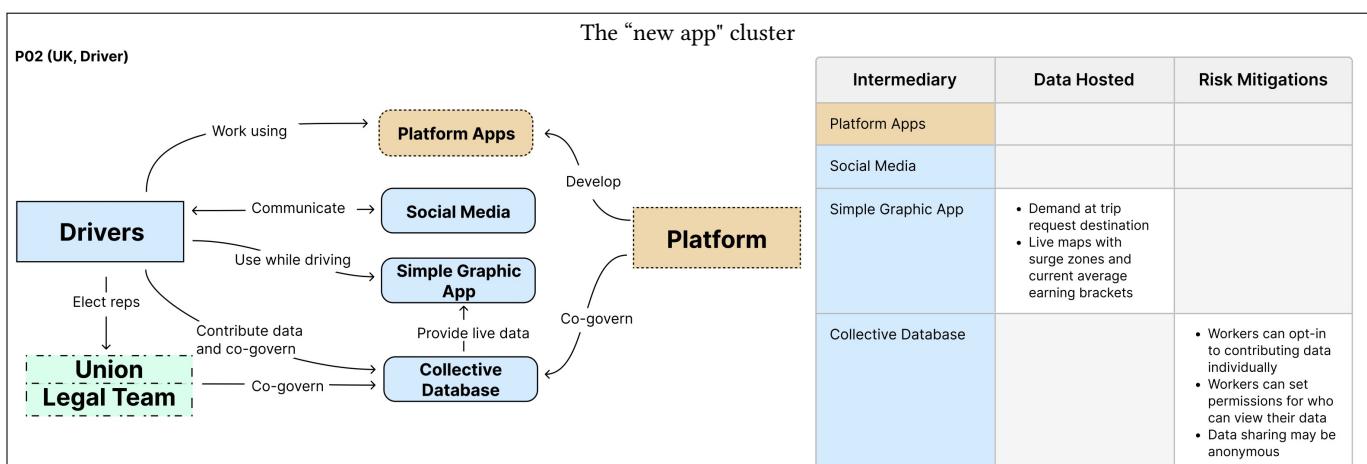
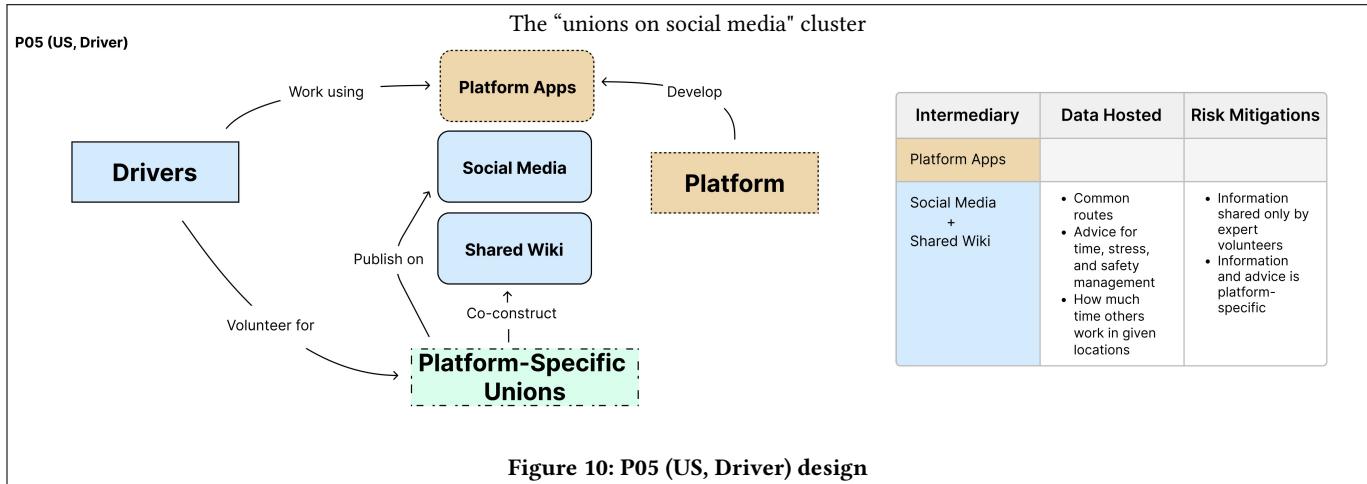


Figure 9: P04 (UK, Courier) design



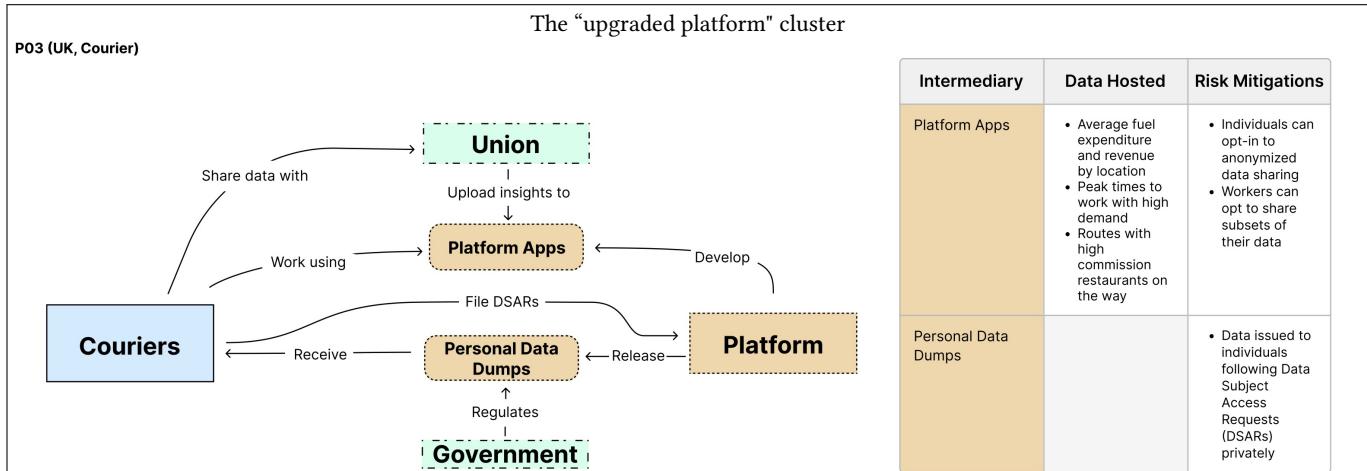


Figure 13: P03 (UK, Courier) design

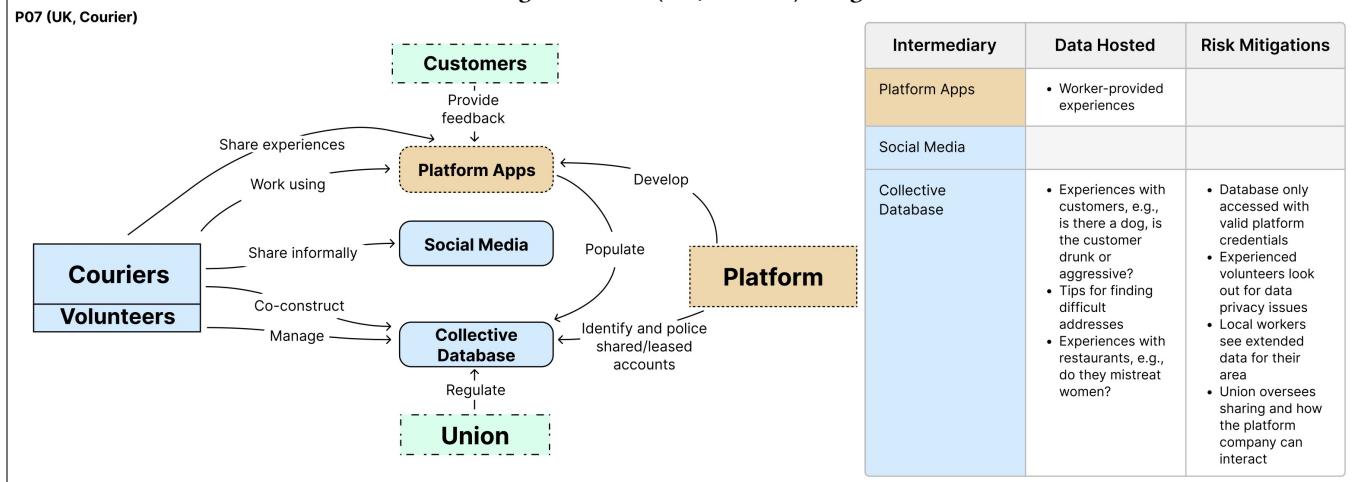


Figure 14: P07 (UK, Courier) design

5 DISCUSSION

The results of our PD sessions with workers provide empirical evidence for the broad plurality, complexity, and subjectivity that data institutions must consider if they are to create meaningful change for workers. Further, setting our work against the backdrop of contemporary demands for HCI to address the social inequalities of capitalism with the self-awareness of also “designing within capitalism,” we hope to advance the discourse by documenting the practical challenges of implementing social computing systems in a hostile environment [55]. These findings and theoretical implications are tied by their mutual necessity in actualizing the vision of alternative institutions which resist platforms in a setting where platforms hold all the cards (or indeed the data).

In this section, we point out the key tensions which emerged among our findings, and situate them within Mouffe and Laclau’s critical theory, which seeks to develop resistance against hegemonic capitalism via “agonistic pluralism” [33].

5.1 Translating Socio-Institutional Plurality Into Usable Data Ontologies

Putting Mouffe and Laclau’s theory to work in overcoming platform hegemony suggests that we need to embrace a constellation of smaller data institutions which honor the relational autonomy defined by workers’ individual working conditions and identities. We found that workers’ visions for collective data governance differed depending on their priorities, needs, and identity. This was evidenced by the variety of institutions that workers generated, and the variable motivations they cited. For example, two workers referenced their marginalized identities (one female, one an immigrant) as the motivation of collective uses of data that might support their safety. These workers suggested using collective data to enforce fair practices and emphasized the need for greater predictability on the platform, while other workers suggested data policies which targeted income optimization. Other workers asserted the equally valid interpretation that sharing data might upset their ability to take advantage of platforms’ idiosyncrasies and make ends-meet in an atomized work environment. Suggesting universal changes to platforms concedes that workers’ individual identity-bound needs cannot be met or addressed. Mouffe and Laclau suggest that supporting multiple, contradictory avenues of resistance is necessary when power is as concentrated as it currently is in platform work.

In practice, workers still must overcome not only information asymmetries with all their harms, but the capital and power asymmetries which reify challenges to adoption, technical overhead, and expertise that come with the task of building sustainable data architectures for grassroots data institutions. These challenges are not posed for platforms funded by endless venture capital, or established data trusts which can lean on well-established institutional supports. As a result, the literature largely ignores them in prescribing design solutions, taking the platform as it is as a starting point.

We found multiple clusters of potential solutions; many of them do not require platform buy-in and would arguably avoid data access asymmetries. These are not necessarily the ideal solutions, but instead, proof that workers are capable of designing data institutions with potential.

The data sharing decisions participants articulated were not merely the result of trade-offs in how much to share and how much to gain, but were informed by a tangle of assumptions about how collective data sharing may work. Mitigation strategies like obfuscating exact payment amounts, sharing data in aggregate form, or delegating governance decisions give a sample of how small changes in institutional design can radically alter willingness for adoption, but it also surfaced the role opaque assumptions play in how participants imagined data governance to work. For instance, several workers feared that data sharing might upset the balance algorithmic management has struck, upsetting earning opportunities for all involved. Accounting for these brings forward a critical challenge for involving workers in data architecture design and a limitation to our study. The folk assumptions about the functionality of algorithms and how they may react to resistance and the decisions workers make may be too inextricably linked to untangle even for workers themselves [29]. On the other hand, the wide plurality of preferences expressed might also be interpreted as a need for technical implementations which can accommodate many different modalities of resistance simultaneously, like the necessity for granular user controls or interoperable data structures.

We also see the complexity and nuance of worker preferences that we encountered as a call for open and highly mutable tools for collective sense-making such that workers can contextualize and unpack the meaning of platform behaviours rather than making assumptions alone. Workers frequently suggested Wikimedia-like designs in which knowledge was contestable, mutable, and open to be expressed in variable formats of knowledge. Such tools could certainly help address not only the gulf in resources for data normalisation, but also in navigating opaque assumptions or influences on data governance decisions.

Furthermore, our participants demonstrated immediate ability to not only interpret the meaning of individual field names and the ability to arrive at metrics by grouping of fields within CSVs data as it was presented in Uber’s DSAR response, emulating data architectures replete with access control structures, risk and threat models as well as joins and transactions between disparate data types based in the requirements of metrics they self-defined. Accordingly, another core consideration in future research should be creating data architectures which might allow for individual mutability, but sufficient interoperability to allow for lightweight ephemeral data ontology construction which might take advantage of the sense-making activities produced via the proposed Wiki-style tools or repositories. Also useful in this regard, would be a means to formally describe data relations, to facilitate the mapping of common priorities or preferences to appropriate data structures.

5.2 Study Limitations and Further Work

The small sample size of our participant pool and open-ended qualitative study design mean that findings of this study are holistic and non-prescriptive. We built a broad general picture of alternative data futures and potential uses for collective data, but not how prominent these concepts are nor how they should be prioritised. A quantitative larger scale study would be required to make such claims.

Approach	Strengths	Weaknesses
Collective advocacy	Does not require new technical infrastructure	Relies on platform co-operation for change
Blockchain	Independent of platforms	Requires new technical infrastructure
Collective wiki	Everyone can contribute Works with off-the shelf technical infrastructure	Requires moderation to maintain Suggested designs rely on platforms releasing more data
Unions on social media	No dependency on platforms Does not require new technical infrastructure	Needs constant maintenance by volunteers Not everyone can contribute
New app	Does not rely on platforms for new features	Requires new technical infrastructure Proposed designs require additional data outputs from platforms
Upgraded platform	Easy transition – platform workers do not need to learn new habits	Requires platforms to do most of the work

Table 2: Strengths of weaknesses of clusters of proposed approaches

In RQ1, we questioned the suitability of PD as a research method for designing alternative data futures. However, in our study design, we implemented only one type of scaffolding for participants using Miro. A comparative evaluation study involving more or less scaffolding than this one could help to further insights on this question.

This study only looked at platform workers' contributions – but our participants identified stakeholders with important supporting roles in their data collective designs, such as customers, union employees, government representatives and legal experts. To build a full picture with more precise and realistic data collective designs, interviewing these other stakeholders would also be important.

With workers at the helm, future research should target each stage in the algorithmic system production pipeline with the same criticism towards its potential to harbor hegemony: from data collection, to training, to deployment and inference. A great place to start is the very beginning: data access. While our participants operated on the premise that they had Uber's omniscient view into worker data, gaining any visibility into data – not to mention the infrastructure to analyze it – can be a Sisyphean task of technical, social and legal obstacles. Perhaps a comparative study of other schemas sourced via subject access requests would be a meaningful analysis to seed a larger repository of counter-data architectures. After all, if workers are to rebuild data institutions in their interest, they'll need the data first.

5.3 A Note on Reification of Inequalities via Platform Work Research

Finally, we reflexively consider how the role of offering participation in short term, one-off engagements in exchange for a fixed reward is unavoidably reminiscent of the platform work we aimed to investigate itself. Placing flyers in community spaces which offer such an arrangement may seem to some users an intrusion of the dynamics of work into community support spaces. Further, research platforms such as Prolific Academic operate in similar models to

gig platforms, providing a venue for such work. The Fairwork Foundation likewise publishes ways in which these platforms should improve to meet minimum acceptable working conditions [21]. In our own work, we emphasized the importance of questioning where platforms begin and end in a broader ecosystem of social media, messaging and forums, as a means to broaden our understanding of how to build resistance, but recruiting in community spaces also risks transgressing such boundaries, making social institutions sites of platforms work themselves. We point to this complication as a further reason for the establishment of small-scale community-run data institutions, which might themselves negotiate research participation as a critical data stewardship activity.

We also recommend partnering closely with advocacy organizations for participatory action research to advocate for workers' autonomy in ways they approve see, for example, Calacci and Pentland's recent work with Shipt workers or Irani and Silberman's reflection after their creation of Turkopticon [9, 27]. This reflection is in part why we advocate against top-down or monolithic solutions which target platforms with specific changes. Instead, it is better to work with advocacy organizations so they may have a say in how they are "being iterated," as Dourish echoes Ahmed's assertion that participation as research subjects is not enough when systemic change in institutional relations is needed [15].

6 CONCLUSION

This paper has explored workers' roles in developing and supporting counter-data architectures. It investigated what it would mean to create institutions which challenge platforms not only as technologies, but as actors situated in the material conditions of labor for platform workers. In doing so, we asked workers what would be necessary to support resistance against platforms, placing them in the role of data analyst and architect. With the ingredients of existing asymmetries, they displayed a keen ability to reconstruct

ontologies and architectures to promote their own independent interpretations of autonomy. Along the way, they found any counter-data architecture needs to be contextual and subjective, yet neutral and fair. It needs all voices to uplift one another, without shading out minority interests in the grassroots. It needs to return meaningful value without incurring disproportionate draws on workers' time. It needs to move quickly enough to keep up with a changing landscape of platforms, but intentionally enough to maintain accuracy and context. It needs to react to the platform, without being solely determined by it. It needs to support the overall autonomy of all workers, without sacrificing the relational autonomy of individuals. Each of these tensions are what workers needed to accommodate in their design for counter-data institutions. The result was plurality, and inherent to plurality is conflict. Indeed, each of these tensions pose challenges to gaining the sufficient momentum and cooperation among atomized platform workers that has held them back to date.

Participants described systems that allow functionalities for worker-mediated voluntary data collection, discourse, and collective sense making. They stressed the need for localization, and granular access control, but also to enable delegation of data control and analysis to collectives, experts, and representatives.

The most related prior work to ours is Zhang et al. from CHI 2022, which placed workers at the center of PD exercises, sketching their algorithmic imaginaries as requirements for more empowering futures [62]. Also, Jia-Jun Li et al., from CHIWork 2022, is similar in its focus on local worker-designed architectures beyond the platform [28]. Unlike most work on algorithmic management, this work contributes a focus on data architecture and social institution design while still drawing on the technique of worker imaginaries. At the same time, we seek to answer other CHI scholarship in critical computing to acknowledge capitalism's influence on design practices. In the vein of [20], we introduce critical theory to motivate a design practice which rejects platform domination as the assumed design environment, instead including it as one of many material conditions of labor which inform worker resistance. Finally, heading Irani and Silberman's reflections on intervening in worker struggles as designers [27], we place workers in the role of data architects to elicit the systems needed to support counter-data architectures like Irani and Silberman's 2013 CHI work [26].

Consensus is not always the path to autonomy. Mouffe and Laclau give us a theory for how to create space for a plurality of conflicting interpretations of resistance, embracing difference as a means to collective power. They consider the greatest challenge for democratizing capitalist power "not how to overcome the we/they relation but how to envisage forms of construction of we/they compatible with a pluralistic order" [44]. In PD, HCI gives us the methods to honor difference and subjectivity, and as a result our participants have given us a start in addressing the necessary challenges to support bottom-up architectures.

REFERENCES

- [1] Ifeoma Ajunwa. 2018. Algorithms at work: productivity monitoring applications and wearable technology as the new data-centric research agenda for employment and labor law. *. Louis ULJ* 63 (2018), 21. ZSCC: 0000041 Publisher: HeinOnline.
- [2] Shaowen Bardzell. 2018. Utopias of Participation: Feminism, Design, and the Futures. *ACM Transactions on Computer-Human Interaction* 25, 1 (Feb. 2018), 1–24. <https://doi.org/10.1145/3127359>
- [3] Oliver Bates, Christian Remy, Callum Nash, and Ben Kirman. 2019. The future of techno-disruption in gig economy workforces: challenging the dialogue with fictional abstracts. In *Proceedings of the Halfway to the Future Symposium 2019*. ACM, Nottingham United Kingdom, 1–4. <https://doi.org/10.1145/3363384.3363476>
- [4] Ana Brandusescu and Jonathan van Geus. 2020. *Shifting Power Through Data Governance*. Technical Report. Mozilla Foundation Data Futures Lab, San Francisco, CA.
- [5] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, communication & society* 20, 1 (2017), 30–44. Publisher: Taylor & Francis.
- [6] Gordon Burtch, Seth Carnahan, and Brad N. Greenwood. 2018. Can You Gig It? An Empirical Examination of the Gig Economy and Entrepreneurial Activity. *Management Science* 64, 12 (Dec. 2018), 5497–5520. <https://doi.org/10.1287/mnsc.2017.2916>
- [7] Susanne Bødker and Morten Kyng. 2018. Participatory Design that Matters—Facing the Big Issues. *ACM Transactions on Computer-Human Interaction* 25, 1 (Feb. 2018), 1–31. <https://doi.org/10.1145/3152421>
- [8] Dan Calacci. 2022. Organizing in the End of Employment: Information Sharing, Data Stewardship, and Digital Workerism. In *2022 Symposium on Human-Computer Interaction for Work*. ACM, Durham NH USA, 1–9. <https://doi.org/10.1145/3533406.3533424>
- [9] Dan Calacci and Alex Pentland. 2022. Bargaining with the Black-Box: Designing and Deploying Worker-Centric Tools to Audit Algorithmic Management. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–24. <https://doi.org/10.1145/3570601>
- [10] Christina J. Colclough. 2020. *Towards Workers' Data Collectives*. Technical Report. It For Change & The Just Net Coalition, Bangalore, Karnataka.
- [11] Craig Dalton and Jim Thatcher. 2014. What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'. *Society and Space* 29 (2014).
- [12] Christopher A Le Dantec and Carl DiSalvo. 2013. Infrastructuring and the formation of publics in participatory design. *Social Studies of Science* 43, 2 (April 2013), 241–264. <https://doi.org/10.1177/0306312712471581>
- [13] Sylvie Delacroix and Neil D Lawrence. 2019. Bottom-up data Trusts: disturbing the 'one size fits all' approach to data governance. *International Data Privacy Law* (Oct. 2019), ipz014. <https://doi.org/10.1093/idpl/ipz014>
- [14] Lynn Dombrowski, Adriana Alvarado Garcia, and Jessica Despard. 2017. Low-Wage Precarious Workers' Sociotechnical Practices Working Towards Addressing Wage Theft. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 4585–4598. <https://doi.org/10.1145/3025453.3025633>
- [15] Paul Dourish, Christopher Lawrence, Tuck Wah Leong, and Greg Wadley. 2020. On Being Iterated: The Affective Demands of Design Participation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–11. <https://doi.org/10.1145/3313831.3376545>
- [16] Veena Dubal. 2017. Wage Slave or Entrepreneur?: Contesting the Dualism of Legal Worker Identities. *California Law Review* 105, 1 (2017). <https://doi.org/10.15779/Z38M84X> Publisher: Berkeley Law Scholarship Repository.
- [17] Veena Dubal. 2021. The New Racial Wage Code. *SSRN Electronic Journal* (2021). <https://doi.org/10.2139/ssrn.3855094>
- [18] Anthony Dunne and Fiona Raby. 2013. *Speculative everything: design, fiction, and social dreaming*. MIT press.
- [19] Lillian Edwards. 2004. The Problem with Privacy. <https://papers.ssrn.com/abstract=1857536>
- [20] Hamid Ekbia and Bonnie Nardi. 2016. Social Inequality and HCI: The View from Political Economy. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 4997–5002. <https://doi.org/10.1145/2858036.2858343>
- [21] Fairwork Foundation. 2022. Prolific (Cloudwork). <https://fair.work/en/ratings/platforms/cloudwork-prolific/?y=2022>
- [22] Mark Graham. 2020. Regulate, replicate, and resist – the conjunctural geographies of platform urbanism. *Urban Geography* 41, 3 (March 2020), 453–457. <https://doi.org/10.1080/02723638.2020.1717028>
- [23] Kathleen Griesbach, Adam Reich, Luke Elliott-Negri, and Ruth Milkman. 2019. Algorithmic Control in Platform Food Delivery Work. *Socius: Sociological Research for a Dynamic World* 5 (Jan. 2019), 237802311987004. <https://doi.org/10.1177/2378023119870041>
- [24] Wendy Hall and Jérôme Pesenti. 2017. Growing the artificial intelligence industry in the UK. *Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy. Part of the Industrial Strategy UK and the Commonwealth* (2017).
- [25] Jack Hardinges and Jared Robert Keller. 2021. What are 'bottom-up' data institutions and how do they empower people? – The ODI. <https://theodi.org/article/what-are-bottom-up-data-institutions-and-how-do-they-empower-people/>
- [26] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 611–620. <https://doi.org/10.1145/2470654.2470742>

- [27] Lilly C. Irani and M. Six Silberman. 2016. Stories We Tell About Labor: Turkopticon and the Trouble with "Design". In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 4573–4586. <https://doi.org/10.1145/2858036.2858592>
- [28] Toby Jia-Jun Li, Lu Yuwen, Jaylexia Clark, Megan Chen, Victor Cox, Meng Jiang, Yang Yang, Tamara Kay, Danielle Wood, and Jay Brockman. 2022. A Bottom-Up End-User Intelligent Assistant Approach to Empower Gig Workers against AI Inequality. *CHIWORK 2022 Symposium on Human-Computer Interaction for Work* (June 2022), 1–10.
- [29] Nadia Karizat, Dan Delmonaco, Motahare Eslami, and Nazanin Andalibi. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produces Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–44. <https://doi.org/10.1145/3476046>
- [30] Becky Kazansky. 2021. 'It depends on your threat model': the anticipatory dimensions of resistance to data-driven surveillance. *Big Data & Society* 8, 1 (Jan. 2021), 205395172098555. <https://doi.org/10.1177/205395172098555>
- [31] Sangmi Kim, Elizabeth Marquis, Rasha Alahmad, Casey S. Pierce, and Lionel P. Robert Jr. 2018. The Impacts of Platform Quality on Gig Workers' Autonomy and Job Satisfaction. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Jersey City NJ USA, 181–184. <https://doi.org/10.1145/3272973.3274050>
- [32] Elsicia Kinder, Mohammad Hossein Jarrahi, and Will Sutherland. 2019. Gig Platforms, Tensions, Alliances and Ecosystems: An Actor-Network Perspective. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–26. <https://doi.org/10.1145/3359314>
- [33] Ernesto Laclau and Chantal Mouffe. 1985. *Hegemony and socialist strategy: towards a radical democratic politics*. Verso, London.
- [34] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (Jan. 2018), 2053951718756684. <https://doi.org/10.1177/2053951718756684>
- [35] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 1603–1612. <https://doi.org/10.1145/2702123.2702548>
- [36] Min Kyung Lee, Ishan Nigam, Angie Zhang, Joel Afriyie, Zhizhen Qin, and Sicun Gao. 2021. Participatory Algorithmic Management: Elicitation Methods for Worker Well-Being Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Virtual Event USA, 715–726. <https://doi.org/10.1145/3461702.3462628>
- [37] Sebastián Lehuedé. 2022. When friction becomes the norm: Antagonism, discourse and planetary data turbulence. *New Media & Society* (July 2022), 146144482211084. <https://doi.org/10.1177/1461444822110844>
- [38] Silvia Lindtner, Shaowen Bardzell, and Jeffrey Bardzell. 2016. Reconstituting the Utopian Vision of Making: HCI After Technosolutionism. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 1390–1402. <https://doi.org/10.1145/2858036.2858506>
- [39] Natasha Lomas. 2021. Dutch court rejects Uber drivers' 'robo-firing' charge but tells Ola to explain algo-deductions. <https://social.techcrunch.com/2021/03/12/dutch-court-rejects-uber-drivers-robo-firing-charge-but-tells-ola-to-explain-algo-deductions/>
- [40] Elizabeth B. Marquis, Sangmi Kim, Rasha Alahmad, Casey S. Pierce, and Lionel P. Robert Jr. 2018. Impacts of Perceived Behavior Control and Emotional Labor on Gig Workers. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Jersey City NJ USA, 241–244. <https://doi.org/10.1145/3272973.3274065>
- [41] Amanda Meng and Carl DiSalvo. 2018. Grassroots resource mobilization through counter-data action. *Big Data & Society* 5, 2 (July 2018), 205395171879686. <https://doi.org/10.1177/2053951718796862>
- [42] Ruth Milkman, Luke Elliott-Negri, Kathleen Griesbach, and Adam Reich. 2021. Gender, Class, and the Gig Economy: The Case of Platform-Based Food Delivery. *Critical Sociology* 47, 3 (2021), 357–372. <https://doi.org/10.1177/0896920520949631> eprint: <https://doi.org/10.1177/0896920520949631>
- [43] Chantal Mouffe (Ed.). 1992. *Dimensions of radical democracy: pluralism, citizenship, community*. Verso, London ; New York.
- [44] Chantal Mouffe. 2005. *On the political*. Routledge, London ; New York. OCLC: ocm56806700.
- [45] Dave Murray-Rust, Alan Davoust, Petros Papapanagiotou, Areti Manataki, Max Van Kleek, Nigel Shadbolt, and Dave Robertson. 2018. Towards Executable Representations of Social Machines. In *Diagrammatic Representation and Inference*, Peter Chapman, Gem Stapleton, Amirouche Moktefi, Sarah Perez-Kriz, and Francesco Bellucci (Eds.). Springer International Publishing, Cham, 765–769.
- [46] Kieron O'hara. 2019. *Data Trusts: Ethics, Architecture and Governance for Trustworthy Data Stewardship*. University of Southampton. <https://doi.org/10.5258/SOTON/WSI-WP001>
- [47] Open Data Institute and Projects by IF. 2022. *Perceptions of "Bottom-up Data Institutions": A report on research findings and key learnings for ODI*. Technical Report. Open Data Institute, London.
- [48] Alex Rosenblat and Luke Stark. 2015. Uber's Drivers: Information Asymmetries and Control in Dynamic Work. *SSRN Electronic Journal* (2015). <https://doi.org/10.2139/ssrn.2686227>
- [49] Herbert Rubin and Irene Rubin. 2012. Qualitative Interviewing (2nd ed.): The Art of Hearing Data. SAGE Publications, Inc., Thousand Oaks, 64–78,223. <https://doi.org/10.4135/9781452226651>
- [50] Anouk Ruhaak. 2020. Data trusts: what are they and how do they work? <https://www.thersa.org/blog/2020/06/data-trusts-protection>
- [51] Bhavani Seetharaman, Joyojeet Pal, and Julie Hui. 2021. Delivery Work and the Experience of Social Isolation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–17. Publisher: ACM New York, NY, USA.
- [52] Brandon Sellers. 2020. Who are delivery drivers: A demographic breakdown of delivery drivers in the U.S. <https://gridwise.io/blog/b2b/who-are-delivery-drivers-a-demographic-breakdown-of-delivery-drivers-in-the-u-s/> Section: B2B.
- [53] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruzen, Angeles Martinez Cuba, Giulia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 667–678. <https://doi.org/10.1145/3531146.3533132>
- [54] John Vines, Rachel Clarke, Peter Wright, John McCarthy, and Patrick Olivier. 2013. Configuring participation: on how we involve people in design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 429–438. <https://doi.org/10.1145/2470654.2470716>
- [55] Christine T. Wolf, Mariam Asad, and Lynn S. Dombrowski. 2022. Designing within Capitalism. In *Designing Interactive Systems Conference*. ACM, Virtual Event Australia, 439–453. <https://doi.org/10.1145/3532106.3533559>
- [56] Christine T. Wolf, Haiyi Zhu, Julia Ballard, Min Kyung Lee, and Jed R. Brubaker. 2018. The Changing Contours of "Participation" in Data-driven, Algorithmic Ecosystems: Challenges, Tactics, and an Agenda. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Jersey City NJ USA, 377–384. <https://doi.org/10.1145/3272973.3273005>
- [57] Alex J Wood, Mark Graham, Vili Lehdonvirta, and Isis Hjorth. 2019. Good Gig, Bad Gig: Autonomy and Algorithmic Control in the Global Gig Economy. *Work, Employment and Society* 33, 1 (Feb. 2019), 56–75. <https://doi.org/10.1177/0950017018785616>
- [58] Alex J Wood and Vili Lehdonvirta. 2021. Antagonism beyond employment: how the 'subordinated agency' of labour platforms generates conflict in the remote gig economy. *Socio-Economic Review* 19, 4 (Nov. 2021), 1369–1396. <https://doi.org/10.1093/ser/mwab016>
- [59] Alex J. Wood, Nicholas Martindale, and Vili Lehdonvirta. 2021. Dynamics of contention in the gig economy: Rage against the platform, customer or state? *New Technology, Work and Employment* (Sept. 2021), ntwe.12216. <https://doi.org/10.1111/ntwe.12216>
- [60] Jamie Woodcock and Mark Graham. 2019. The gig economy. *A critical introduction*. Cambridge: Polity (2019).
- [61] Zheng Yao, Silas Weden, Lea Emerlyn, Haiyi Zhu, and Robert E Kraut. 2021. Together But Alone: Atomization and Peer Support among Gig Workers. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29. Publisher: ACM New York, NY, USA.
- [62] Angie Zhang, Alexander Boltz, Chun Wei Wang, and Min Kyung Lee. 2022. Algorithmic Management Reimagined For Workers and By Workers: Centering Worker Well-Being in Gig Work. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–20. <https://doi.org/10.1145/3491102.3501866>

ACKNOWLEDGMENTS

This research was supported by the Oxford Martin School Programme for Ethical Web and Data Architectures in the Age of AI at the University of Oxford.

A CODEBOOK SUMMARY

Our full codebook is comprised of fifteen top-level codes with variable numbers of lower-level codes and can be found in the supplementary materials of this paper.

The most important top-level codes can be categorized into three groups: 1) participants' backgrounds and demographics, 2) participants' ability to understand data, their motivations for data collectivization and their personal data privacy preferences, and 3) designs for alternative institutions that participants put forward.

The first group includes codes used to describe the participants' backgrounds and demographics. We coded these details under "Participant background," which included aspects like hours worked per week, the platforms they use, their previous work experience, union involvement and work habits.

The second group of codes marks sections related to participants' ability to understand data, their motivations for data collectivization and their personal data privacy preferences. These include "Collectivization sentiment," when workers expressed positive or negative feelings about collectivizing data, to be used in tandem with "Collectivization motivations," which includes a variety of reasons why workers were interested in collectivizing data (from algorithmic transparency or just altruism towards other workers to income optimization, safety and building legal appeals to platforms). Finally, "Data sharing & privacy," was used for instances and sentiments expressed by workers about willingness to share data or concerns about sharing data with other workers. We categorized these sentiments into closed, mixed, open and neutral. "Individual governance," was used to flag instances when participants described their own (potential) involvement in data collective governance decisions.

The third main group of codes was used to capture the designs for alternative institutions that participants put forward. These include "Data access methods," for example "new app," "blockchain," and "Wikimedia-style," and other forms of communication like "data dumps," "informal," and "unions." "Collective data uses," was used to catalogue which stakeholder should make use of collectivized data, this ranged from lawyers and platforms to governments and unions. While "Subgroups" was used to capture when participants pointed out groups (e.g. ethnicity, drivers vs. cyclists, location-based) within workers who might operate differently within their imagined institutions, "Other stakeholders," was used when participants described stakeholders beyond platforms and workers involved in their imagined institutions. Finally, "Risks," was used to identify potential downsides or threats to institutions which participants considered, including mitigation strategies which they offered to pre-emptively address these.

B PARTICIPANT DATA LITERACY

Instead of asking participants to self-report their data literacy level (which would be inconsistent across participants and potentially unreliable) or asking all participants to complete a pre-assessment (which would be overly time consuming for the participants), we decided to estimate data literacy levels from participant contributions observed during the interviews. Those who shared specific experiences of using data in their work were rated with "High" data literacy. Those who had relevant ideas and understanding, but no experience-based examples were rated "Medium". The remaining participants were rated "Low". Since the level assignments could be subjective, we have included specific quotes from the interviews supporting our assigned data literacy levels below, but readers who disagree with this categorization should interpret found trends relating to data literacy accordingly.

P01 (data literacy = medium).

- "I was one of the founding members of a, you know, now defunct organization that was going to advocate for the rights of for hire

transportation workers and small business owners"

- "personally, I'm not a huge privacy guy cause I think that if you're stupid enough to think that anything you do on your phone is private... Just so to every extent is like... doesn't make sense, especially now that everything is a data mine. Even buying a cup of coffee, there's tracking and selling their data. I think people's personal information should be kept secret to the extent that they want to."

P02 (data literacy = medium).

- "I've tried that with other drivers where we could compare what was on our screen and on their screen. [...] I don't know if Uber is catering a special screen just for me. You know what I mean? Because I have no way to compare with anything, so I have to rely on the information that Uber is providing me as truthful, which sometimes is difficult."
- "I don't accept those jobs because some of them, especially long business ones, they are paying way below what they should. So I don't accept, but I know that for some drivers any job on their screen they accept. I think there are less and less drivers like that."
- "Like a solid chunk of data were treated, naturally, and you would see that visually because if you are using one of the apps you don't have much time to, kind of, move around too much or to have much more detailed data, so [it] has to be more graphical kind of stuff. And an app is great for that."

P03 (data literacy = high).

- "All you're doing [by pooling collective data], really, is increasing the size of your dataset"
- "I've been doing this a couple years. I've done this already, so I did a little... I did a little bit of analytics myself and worked out, sort of, maximum revenue from certain locations for least fuel expenditure..."
- "The more you share your, sort of, assessments from a large dataset, the more people are gonna, you know, modify their behavior accordingly. Which you don't wanna do! You know, the reality is you're looking to obtain an advantage over other drivers. And if they all start working what you've worked out to be the peak time, then it's not gonna be the peak time for very long. Obviously as well, there's... you'd need to make sure the data is completely anonymized, and like I say, the usual data protection considerations."

P04 (data literacy = high).

- "They collect banking information for security purposes. [...] It's to verify that that one person is not running as four people."
- "We don't really use software for it. We just do it on the fly. [...] Yeah, though it's not always 100%. [...] There is software out there to do it, but it's not open source and it's a bit... concerning to put all your information into it."
- "That's actually an interesting one, because we've [in a couriers' group chat on social media] previously done that [pooling data together]. And for two categories of customers, so customers who tip good (so we're talking about a couple of quid extra). And customers who are... well, just a bit of a ****."

P05 (data literacy = low).

- "[Interviewer: what do you think you can learn from your own data, if you have access to all the data that the platform is collecting about you?] OK, I think it will help to make my work easier, and

accessibility to the people I'm working with, and proper records, proper records, proper records, keeping of records on everything, that the things that [we] do in a typical day, so that if I'm to locate it or if I'm to have a review on that the next day, it will be really difficult for me. So I think it will also help me to be more organized."

- "This will help me to segment my data from the data of other drivers without any much difficulty because I have a foresight of all my data and I have a foresight of what I'm working with. So if I'm to... if there [were] to be a mix up of data or something, I will quickly know how to differentiate each data because I have a particular structure on, or have a particular style of database that I work with, so that will make work very easy for me so..."

P06 (data literacy = medium).

- "They know my work location because when we join the platform, they ask us which will be our central location where we would like to start from, majorly. Some people give home location as work as well because they are very close to city center, like cycle riders."
- "The reason Uber does not reveal specific customers' feedback to us is some drivers might go back to their house and start a fight. Like you did, you give me a bad rating or you know, feedback. [...] This is one lingering and very, very long pending issue that Uber decides our rating and grading on and that affects our payments. But there is no chance that customer or the drivers will know who has done this because the feedback comes in about 7 to 10 days and by that time driver cannot go back..."
- "that is a reality that people are running multiple platforms on their phones. And this is one of the reasons the deliveries get delayed is... they pick up the order on Uber. Then from the same restaurant, they pick up another order on Just Eat or Deliveroo. A customer could be going in X direction, the delivery order could be in the Y direction, so he makes a call who to deliver first. So anyway, he's carrying four orders so he wouldn't want the other drivers to notice that he's actually making more money in a trip."
- "McDonald's is infamous for wait times. Every time we pick up an order from McDonald's, it's between 25 minutes to 40 minutes to make £5."

P07 (data literacy = medium).

- "I suppose probably from a driver's perspective, how much, sort of, swapping of the accounts between a different driver? Because that seems to happen quite a lot where someone will sell their account for the day. And you don't know who's actually supposed to be the driver."
- "I suppose the risk could be that if you showed too much information, people would not want to work. Or. They would perhaps not take the work. Or sort of leave it for someone else to do. See if we all left with the rubbish jobs, which wouldn't be fun."
- "You couldn't, sort of, know what sort of privacy each person wants. Cause if they're using the apps, they're pretty, sort of, tech savvy anyway, they've got a bit of an idea what they're doing. You might have to show them, obviously. I don't know if they would probably go down the world of, like, letters or anything. That's probably too costly."

P08 (data literacy = high).

- "As a driver, a lot of my interest comes from the algorithms of the company themselves and how they use this information to decide

not just which offers are sent to us, but the pay associated with those."

- "No, this doesn't surprise me, and I'm sure there's much more data they're tracking in the form of cookies and advertising revenue that aren't present on these lists."
- "The first question I would ask is how they are scoring us internally... Which metrics they are using to score us with internally, that they don't share with us. I've actually been a beta tester with DoorDash for many years. [...] But I know that they are using those internal metrics to program the algorithm to determine who gets what offers."
- "I think Para tried to pivot towards a prediction-based model, likely using information that they're screen capturing from our offer screens and earnings screens at the end, which they now openly admit, 'we are breaking TOS by doing'. I've never used it, but at the same time you've got that other app to accept or decline. I think it used to be called Driver Utility Helper. Which is just using the accessibility to automate the tasks on your app automatically. Decline orders that don't meet a criteria. One click declines so you don't have to go through other whole list of reasons and things like that. That one's focused more on usability, whereas Para was focused on revealing hidden pay data."

P09 (data literacy = medium).

- "We knew we were being paid before they changed the pay scale. And then they changed it to this black box algorithm, but anyways. They've over hired so much and, when you're brand new, you're supposed to get your first ten orders like hand fed to you, so it like bypasses everybody, regardless of their stature. But now we are seeing people doing orders that we know for a fact, because they have told us, that they have horrible ratings, horrible reviews, but yet they have orders while we're sitting in our cars. When we have perfect stats."
- "They [customers] are not required to give a reason. So suddenly we have this three star rating and there's no feedback or no growth opportunity. So we don't know what we did, if anything, to fight it. You know, be like, no, that's not it. She got her can of, you know, soup or whatever. So there's no way... they don't make them do that, and then it's just it's just all the, I call it the Wizard of Oz effect, because it's always there's nothing to see here. Don't look behind the curtain. You know, they don't want you to know anything."
- "they want you to keep at least a four point... I think it's a 4.9 rating. And if you drop below that, they'll give you time to get it back up. But if you don't, they let you go. But you know, you might not know that."
- "Like we have a Facebook page and we can be like, 'hey, prepaid orders aren't processing' and somebody will post that. That way everybody else knows, because maybe nobody else has ran across it yet."

P10 (data literacy = medium).

- "Yeah, I mean we would know, like, if we have all the data of each other, we would be insured that we are being paid fairly. And we would reduce abuse and the cheating by the customers. Well, we've had a lot of experiences of cheating... [some customers call out Ubers and then cancel them before arrival. Or contact support and say the driver didn't do their job to get a refund]"
- "[Interviewer: Would you want to share your data with them

[Uber drivers like you?] If they're from different areas, and if they are in completely different locations, then sure."

P11 (data literacy = high).

- "It's a pretty short list, right? Like I can think of 1000 other things. [...] Well, I guess there is... was that messages to and from riders that probably includes any phone calls. You know, because there's audio recording... you have to allow audio recording to use the app, so they are recording audio at some point. I think that's something that a lot of drivers are curious about."
- "They track data about, like, how you use the app and what else is open on your phone. [...] they will know if you're playing Candy Crush."
- "Well, and I mean, literally everything runs off of like AWS or whatever [...] I ran a firewall on my phone for a while and it was

all like that's... hundreds of... hundreds and hundreds of contacts that are all stored at Amazon."

C PARTICIPANT DATA SHARING PREFERENCES

Participant choices in the Miro board exercise on sorting types of data into buckets of whom they would be willing to share it with are summarized in Table 3.

D WHITEBOARD EXERCISES

In the following pages, Figure 15, Figure 16, Figure 17, Figure 18, and Figure 19, provide examples of both incomplete and complete exercises performed by the participants.

	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11
Account Support Tickets	3	6	5	5	1	2	6	2	4	5	2
Trip Data	1	7	1	1	6	6	1	6	4	5	1
Ratings	5	6	5	2	6	1	1	1	1	5	5
Trusted Contacts	3	6	5	5	4	5	6	5	5	5	5
Payment Information	3	1	5	3	2	5	5	5	4	5	6
Dispatches Offered and Accepted	1	1	1	1	6	6	5	6	4	5	2
Cancellations	1	1	1	1	1	1	1	1	4	2	3
Saved Locations	5	6	5	5	2	5	2	5	1	5	5
Online/Offline	5	1	5	1	6	4	6	4	4	5	2
Device Data	2	3	5	1	6	4	6	4	1	5	6
Bans Restrictions	2	6	5	5	1	1	6	1	1	5	3
Messages	2	6	5	5	6	5	5	5	4	5	3
Profile	1	6	5	5	6	1	3	1	1	5	3
Performance Awards	5	6	5	6	1	1	1	1	1	5	6
Safety Complaints	1	6	5	5	6	1	6	1	1	2	6

Table 3: Participants' Data Sharing Permissions Decisions

- 1) Share with all other drivers 2) Share with managers of a collective 3) Share with a self-defined subgroup of drivers
 4) Share, but exclude a defined group 5) Don't share with anyone 6) Did not categorize 7) More than one category

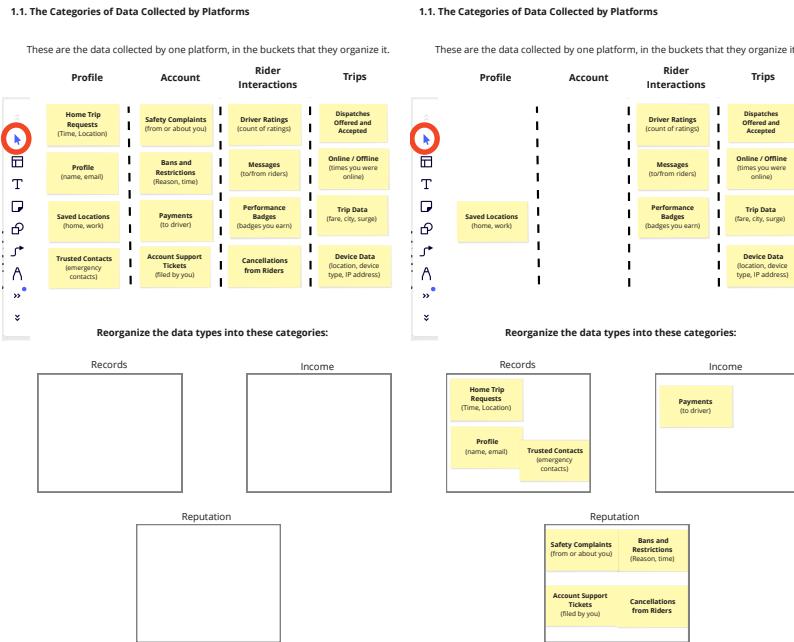


Figure 15: Data categorizing exercise on Miro, before and after completion

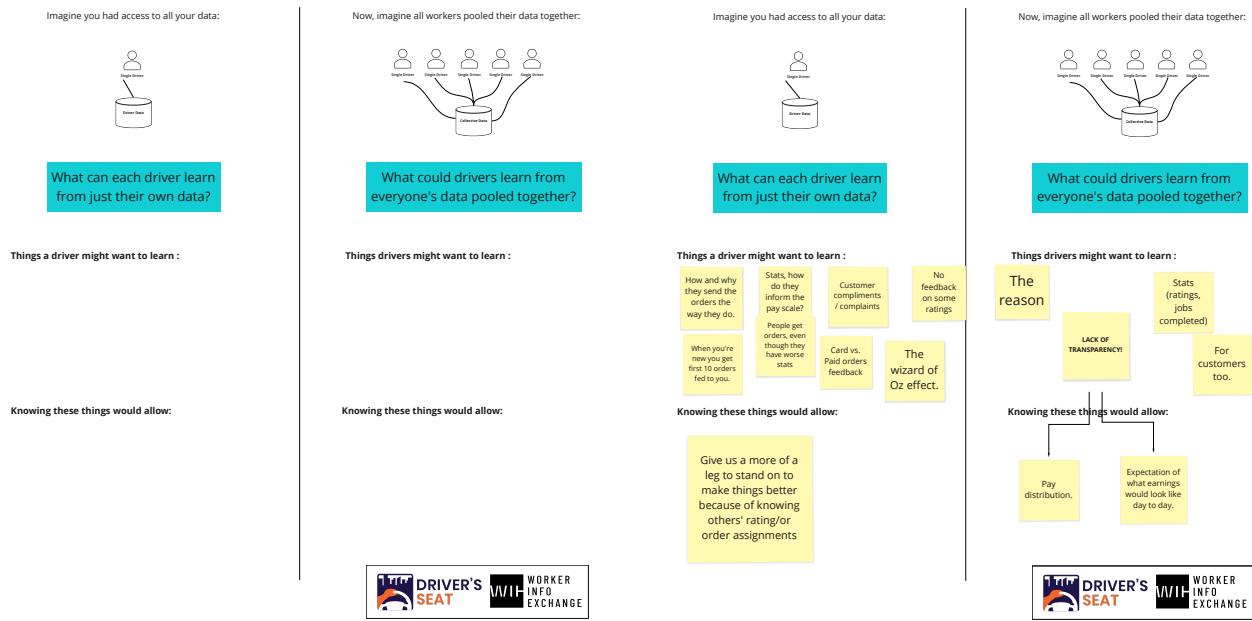


Figure 16: Brainstorming about data uses exercise on Miro, before and after completion

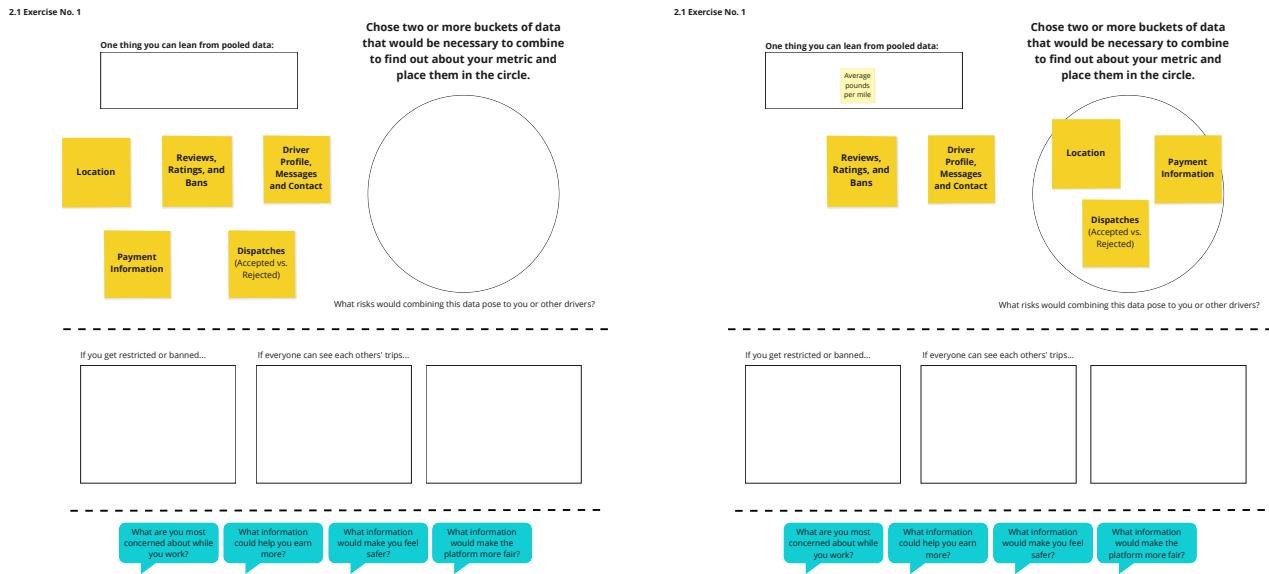


Figure 17: Data combination exercise on Miro, before and after completion

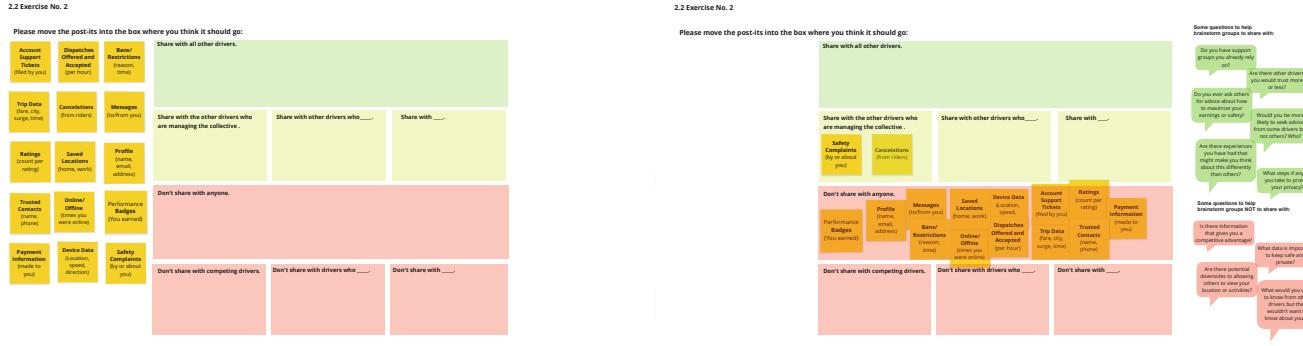
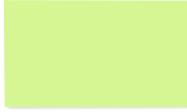


Figure 18: Data sharing card sort exercise on Miro, before and after completion

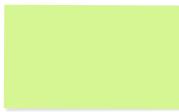
Who should get to make these decisions?



Who should get to make these decisions?



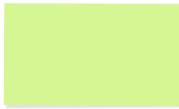
Would you want to participate in these kinds of decisions? How?



Would you want to participate in these kinds of decisions? How?



Would you volunteer your time to help manage this kind of information for the whole collective, or for a smaller group?



Would you volunteer your time to help manage this kind of information for the whole collective, or for a smaller group?



How would you like to access this information?



How would you like to access this information?

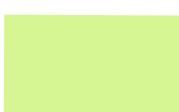


Figure 19: Reflection exercise on Miro, before and after completion

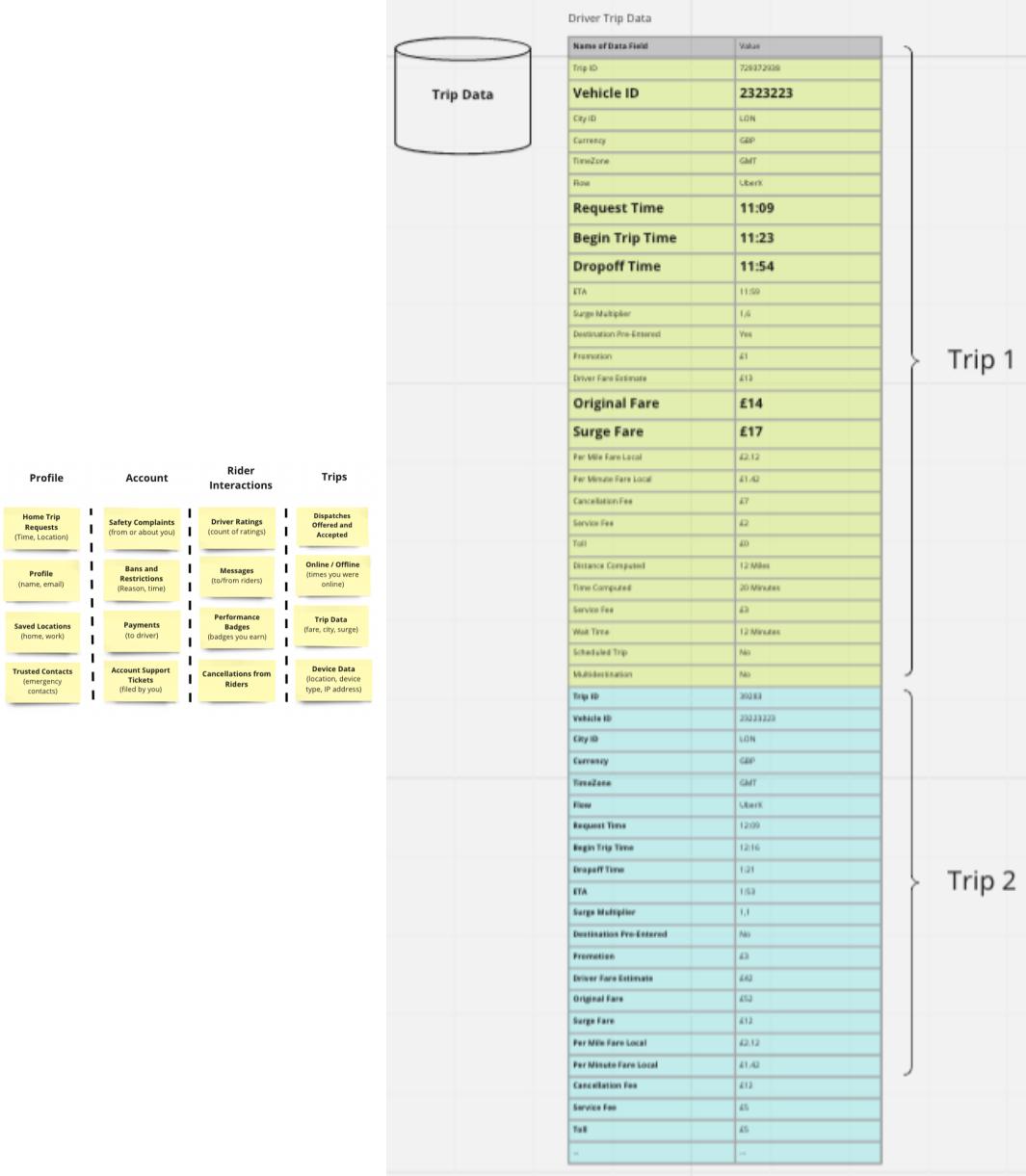


Figure 20



Figure 21: All the representations of the Subject Access Response Schema we provided throughout our exercises including exercise 1, which represents the CSV files and a detailed view with example data and exercise 3, which represents each table with fields as smaller leaves.