Data & Policy (2020), 2: e20
doi:10.1017/dap.2020.21

RESEARCH ARTICLE

Trusts, co-ops, and crowd workers: Could we include crowd data workers as stakeholders in data trust design?

Richard C. Gomer[1,*] and Elena Simperl[2]

[1]School of Electronics and Computer Science University of Southampton, Southampton, United Kingdom
[2]Department of Informatics Kings College, London, United Kingdom
*Corresponding author. E-mail: r.gomer@soton.ac.uk

Received: 02 June 2020; **Revised:** 10 November 2020; **Accepted:** 13 November 2020

Key words: data co-op; data trust; crowd sourcing; crowd work

Abstract

Data trusts have been proposed as a mechanism through which data can be more readily exploited for a variety of aims, including economic development and social-benefit goals such as medical research or policy-making. Data trusts, and similar data governance mechanisms such as data co-ops, aim to facilitate the use and re-use of datasets across organizational boundaries and, in the process, to protect the interests of stakeholders such as data subjects. However, the current discourse on data trusts does not acknowledge another common stakeholder in the data value chain—the crowd workers who are employed to collect, validate, curate, and transform data. In this paper, we report on a preliminary qualitative investigation into how crowd data workers themselves feel datasets should be used and governed. We find that while overall remuneration is important to those workers, they also value public-benefit data use but have reservations about delayed remuneration and the trustworthiness of both administrative processes and the crowd itself. We discuss the implications of our findings for how data trusts could be designed, and how data trusts could be used to give crowd workers a more enduring stake in the product of their work.

Policy Significance Statement

Data trusts are of interest to policy-makers as a means of governing data that is involved in the policy-making process. To date, the discussion of data trusts has centered on their legal form, and on how they can address the interests of groups such as data users and data subjects. However, the poor employments conditions of many crowd workers pose a challenge to the ethical and legitimate use of crowd-worked data. Including the interests and wishes of crowd workers in the design of data trusts could improve the legitimacy and sustainability of using crowd-worked data, and will therefore be of interest to policy-makers who are looking to establish data trusts that do, or might, incorporate crowd-worked data.

1. Introduction

Datasets that are of economic and social value often have complex provenance, arising from a sequence of actions spanning multiple stakeholders. Moreover, once organizations hold datasets, they are frequently not available to other organizations that might be able to put them to productive use. As a result, there is a growing interest in data governance arrangements such as data trusts that can simplify and facilitate data

use and re-use, while balancing the interests of different data stakeholders such as the data producers, potential data users, and the data subjects themselves.

Data trusts are legal structures that hold data on behalf of beneficiaries—such as the public, or a community of crowd workers—and license (or otherwise exploit) the data for the benefit of those beneficiaries; much like a trust might invest a sum of money on behalf of a minor, or preserve and exhibit historical artifacts for the benefit of the public at large. Data trusts have been proposed as a means of increasing the availability of economically or scientifically valuable datasets (Hall and Pesenti, 2017; House of Lords, 2018) or allowing data subjects to disrupt current "feudal" approaches to data governance (Delacroix and Lawrence, 2018).

To date, the discourse and practice around data trusts and similar structures have largely ignored the significant role that manual human labor plays in creating and curating many datasets in the form of crowd-sourcing and crowd work. Crowdsourced "data work" can take various forms, from volunteer projects, such as OpenStreetMap or Zooniverse, where contributors give their time for free to a cause that they feel is important, to paid "microwork" that rewards contributors with small payments in return for completing tasks, like labeling or transcription, through platforms such as Amazon Mechanical Turk.

Crowd microwork provides flexible and accessible work to a wide range of people by allowing them to work remotely over the Web, but concerns abound that it is often unfair or exploitative (Ettlinger, 2016; Fieseler et al., 2017). Ørting et al. (2019) note that, in the context of using crowd workers to create medical datasets, "ethical questions regarding worker compensation, image content and patient privacy are rarely discussed, but seem crucial to address." Furthermore, the microwork format that structures most paid crowd work leaves crowd workers with little stake in the output of their labor, or even any knowledge of how it contributes to a broader goal.

Drawing on stakeholder theory, we argue that crowd workers are a recipient stakeholder group who are affected, negatively or positively, by the ways in which datasets are curated and exploited, and that there is scope for novel governance models such as data trusts or data co-ops to attend to the needs of crowd workers. Miles (Miles, 2017) observes that:

> "it is by defining what is and what is not a stakeholder that we create the reality of whose interests are, and are not, attended to and, in turn, discriminate what is, and is not, empirically tested by academics, attended to by managers or regulated in practice."

In this paper, we deliberately focus on the needs and wishes of crowd workers, and attempt to situate them within the emerging landscape of data trusts. We present the results of our preliminary enquiry into crowd workers as a data stakeholder that could—and we'd argue should—be represented in the design of data governance vehicles like data trusts and co-ops.

We present the results of two studies: The first, a survey of extant and proposed data trusts and data co-ops, identifies salient features of their governance arrangements and stakeholders. The second, a two-phase survey of crowd workers that aims to understand their perceptions of alternative governance arrangements, and factors that might affect the acceptability of those arrangements.

This topic should be of interest to policy-makers for three reasons. First, because where data are used to inform policy, having a robust and ethical data handling and governance process is important to the legitimacy and sustainability of the policy-making process. Second, because the non-commercial public-benefit nature of the policy-making process makes it an ideal candidate to explore innovations that promote justice and equity, as opposed to commercial arrangements where financial concerns are king. And third, because—of course—the outcome of policy-making is, frequently, policy that directly concerns the governance and exploitation of data, and which aims to balance the rights of different data stakeholders that are involved in the policy-making process.

2. Background

Data microwork raises many of the same ethical questions as "gig economy" work more broadly, such as employee rights and exploitation (Fort et al., 2011). The most obvious value in exploring the stake of

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

crowd workers within data governance models, then, is largely in addressing those questions and improving outcomes for workers. There are compelling arguments, beyond ethics, for treating a work force fairly and—even within a dominant system of free-market capitalism—most countries have, albeit to varying degrees, adopted worker protection rules that speak to the social value of secure and rewarding employment. Economically, there are arguments that inequality can itself be inefficient (Borooah et al., 1995) and prior research into co-operative models of employment in non-digital contexts (e.g., Craig and Pencavel, 1995) suggests that they can be advantageous in terms of efficiency, producing a higher output for a given capitalization than their non-co-operative equivalents.

There are three reasons that compel us to consider data trusts as a means of disrupting current crowd work practices.

First, because the datasets that a data trust governs are a discrete capital resource (Sadowski, 2019). Practically, granting crowd workers a stake in a dataset is potentially a pragmatic means of recognizing their work and improving their remuneration via an ongoing return. Not all forms of crowd work produce such a delineated output to which a stake could be attached.

Second, that as O'Hara (2019) notes, a data trust should provide "ethical, architectural and governance support for trustworthy data processing." In the case of crowd-worked data, an ethical approach must surely consider the rights and conditions of data workers themselves.

Finally, because the inequality currently inherent in the crowdwork economy may well be economically inefficient (Gregg et al., 1994), and hence the exploitative nature of crowd work may be hindering the value of human-in-the-loop data processing rather than helping it. Improving conditions for crowd data workers may—contrary to initial impressions—be good for business.

2.1. Crowd (data) work

Crowd work is a form of employment that utilizes distributed human workers to perform tasks through the Web, often for purposes such as training machine-learning models (Wortman Vaughan, 2018), information seeking, or content classification (Gadiraju et al., 2014). Tasks are often structured as very small units of work—"microtasks"—that take only seconds or minutes to complete (Dawson and Bynghall, 2012). Because tasks are completed over the Web, crowd work gives access to a very diverse, geographically distributed workforce (e.g., Ross et al., 2010).

The affordances of the Web mean that many of the tasks carried out by crowd workers can be categorized as what might be termed "data work"—finding, creating, or curating data. The results of crowd work then are frequently datasets, and human computation now forms an important part of economically and socially valuable data value chains (Cavanillas et al., 2016).

Although the term "crowd" includes individuals who engage in projects through voluntary means—for instance in citizen science—in this work, we're primarily interested in crowd *workers*, that is, those whose relationship to the other parties in the activity is one of employment for financial reward. We also limit our inquiry to crowd workers who are engaged for their labor, and not who are themselves data subjects. We recognize, though, that the latter case—in which personal data about crowd workers is incorporated into datasets—would be worthy of study in its own right.

Partly as a result of the geographic (and hence jurisdictional) diversity of the crowd, and partly as a result of the economic inequality between task requesters and workers, crowd working has been criticized as exploitative (Busarovs, 2013; Ettlinger, 2016) and as undermining established employment laws and workers rights (Prassl and Risak, 2016).

Fieseler et al. (2017) explore the fairness of on-demand crowd work in three-way employer–worker–platform relationships. They find that crowd workers report distributive and procedural unfairness, as well as what they term "interactional fairness" which "refers to the quality of the interpersonal treatment employees receive from authorities."

Technical interventions, such as Dynamo (Salehi et al., 2015) have been proposed to help catalyse improvement in crowd workers' employment rights by supporting workers to take collective action, and Stanford's "Fair Work" project (Whiting et al., 2019) has developed a tool that aims to make crowd work

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

fairer by allowing requesters to automatically adjust payments to $15 per hour. Gaikwad et al.'s "Daemo" marketplace (Gaikwad et al., 2017) demonstrated two design interventions— "Boomerang" and "Prototype tasks"—that improve worker–requester trust by addressing two common source of trust breakdown—flawed reputation systems and poorly designed tasks—in crowd microwork platforms.

Kost et al. (Kost et al., 2018) consider how "meaningfulness" is constructed by crowd workers, and conclude that meaningfulness is derived from the work's perceived impact on the self and others. The finding that meaning can be derived from work's impact on others is notable, as although crowd workers in contemporary paid microwork typically lack the information necessary to determine that impact, it suggests that social value, and the broader structures and relationships in which crowd workers are embedded, are important factors to explore.

2.2. Data trusts and data co-ops

Data trusts are a novel legal and organizational structure for governing access and use of datasets (Hardinges, 2018; O'Hara, 2019). Just as a legal trust might (in jurisdictions that have such a concept) be established to invest a sum of money in the interest of a defined beneficiary, so a data trust can be established to govern access to and use of datasets on behalf of a defined group of beneficiaries, the trustees of the data trust bound by a strict fiduciary responsibility to act in the interests of the beneficiaries. The motives and structure of data trusts are similar to "data co-ops" in so much as they create a pooled data resource that can be exploited for the benefit of members, although the legal arrangements for trusts and co-operatives may be different, and—unlike a co-op—a trust might be operated for the benefit of beneficiaries who are themselves not directly involved in the governance process. Although data trusts have similarities with the concept of Trusts that exist in some legal jurisdictions, authors such as O'Hara (O'Hara, 2019) point out that those legal trust structures may not be the most appropriate legal vehicle for creating data trusts, and others such as Stalla-Bourdillon et al. (2019) suggest other legal vehicles such as the Channel Islands' "Foundation" entities. Given the range of potential legal arrangements that could be used to implement data trusts, they are now referred to by many authors using the umbrella term "Data Institutions."

The motivations to establish data trusts can be diverse; for instance, the United Kingdom's House of Lords report into AI (House of Lords, 2018) considers data trusts as a means of unlocking the value of closed datasets by simplifying how access to them is negotiated—data trusts are offered as a repeatable and scalable way of governing datasets that cannot (for commercial or other reasons) be made wholly Open. Data trusts can also change the dynamics of data governance; for instance, Delacroix and Lawrence (2018) propose data trusts as a means of inverting what they describe as the current "feudal" system of personal data governance. For Delacroix and Lawrence, data trusts offer the possibility of pooling personal data (an asset that is more valuable in aggregate than as the sum of its constituent bytes) but (through a fiduciary responsibility) retaining the rights to self-determination that data subjects[1] are now granted in many jurisdictions.

In the case of crowd-produced datasets a data trust might encourage access to, and exploitation of, a particular crowd-worked dataset, while also seeking to maximize benefits to the workers who created or curated it; possibly also taking into account the stakes of other groups such as data subjects, or the rights-holders of any constituent datasets.

The Open Data Institute has undertaken several pilot projects considering how data trusts could be established in various domains (Hardinges et al., 2019). In particular, the ODI pilots show the importance of a design phase in the data trust life-cycle. Considering data trusts as designed artifacts is helpful, as it makes explicit the need to consider stakeholder needs and resolve tensions; and motivates our own work

[1] In EU (and other) data protection law, a data subject is an identifiable individual to whom some data are, or could be, linked. Being the subject of personal data grants (in most jurisdictions) a set of rights that are independent of any notion of "ownership." Data subjects may be "data providers," as in the case of people who contribute their own medical records to a co-op; but they are not necessarily so—for instance, medical records could be provided to a trust by a healthcare organisation.

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

as an initial exploration of the interests of data workers as a—to date, largely invisible—stakeholder group that could, *should* be considered in the data trust design process.

Data trusts have some similarities with other data governance models, in particular data "co-ops." As we'll see later, what exactly constitutes a data co-op in practice is hard to define, many are not worker-owned organizations as co-operatives are commonly understood, but broadly data co-ops are organizations that pool data for use by multiple data consumers.

In the UK, NESTA (Borkin, 2019) has proposed a number of ways that "platform co-operatives" could operate, including "data consortia platforms" in which "a mutual organisation is formed to manage the data on behalf of its members, who have both democratic control and an equitable share in its profits."

It's noteworthy that there is a great deal of discussion about how data trusts might operate and the legal form that they might take (see e.g., Hardinges, 2020); and, at the time of writing, relatively few examples of data trusts (or comparable structures) operating in practice. That diversity is a challenge, insofar as it complicates any attempt to define the object of study or to describe concrete implications for practice, but also indicative of the timeliness of considering a diverse range of potential data trust stakeholders—we are collectively still very much at the design or requirements-gathering stage for what could become important structures in data policy-making and in our data-rich society more generally.

2.3. Fairness

We consider that the question of what stake crowd workers have in the fruits of their labor is (as other questions about crowd worker rights and rewards) fundamentally a question of fairness.

Fairness is a concept that's received considerable attention, but which is conceptualized differently across disciplines. Universally, though, fairness is considered to be similar to—and often synonymous with—justice. In economics, fairness tends to be defined in objective terms, and (for instance) Homans, who developed the field of social exchange theory, defined what he called "distributive justice," as being that "a man in an exchange relation with another will expect that the rewards of each man be proportional to his costs" (Homans, 1961, p. 75).

In practice, outside the idealized scenarios of economics, what's fair is contested. Thus fairness is also political. As a political concept, fairness is harder to define in such objective terms as Homans applies in economics, and disagreements over what constitutes fairness are at the heart of much historical and contemporary political discourse. Marxist, Libertarian, and Ordoliberal perspectives on what is fair in a given scenario might reasonably be expected to differ.

One influential account of fairness from political philosophy is proposed by Rawls (1985). He proposes a thought experiment in which members of a society agree on how their society will function from the so-called "original position"—that is, before they know about factors that affect their position within that society such as wealth, gender, ethnicity, or background. He argues that decisions made from such a position would be fair and that using the original position as a thought experiment can be a useful exercise in considering the fairness of a scenario.

The scope for fairness to incorporate economic and social factors, as well as the relative intuitiveness of it, led us to adopt fairness as a theoretical underpinning for our own enquiry. We framed our second study —the questionnaires—in terms of fairness, and have adopted it as a lens through which to analyze the qualitative data that we gathered.

2.4. Stakeholders in crowd data work

Finally, it's worth considering how crowd workers relate to other stakeholders within data trusts. Figure 1 shows a conceptual schematic of a typical data trust arrangement, based on Open Data Institute (2019a), to which we have added The Crowd for discussion purposes.

A data provider, which has a dataset, entrusts it to a data trust with a set of permissions. Data users then ask the data trust for access to that dataset, and the Data Trust grants access to the data users. Not shown are "data subjects," the people to whom a dataset relates, who are already recognized in many jurisdictions as important stakeholders. Data subjects could be data providers (as in the case of "Bottom-up data trusts"

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

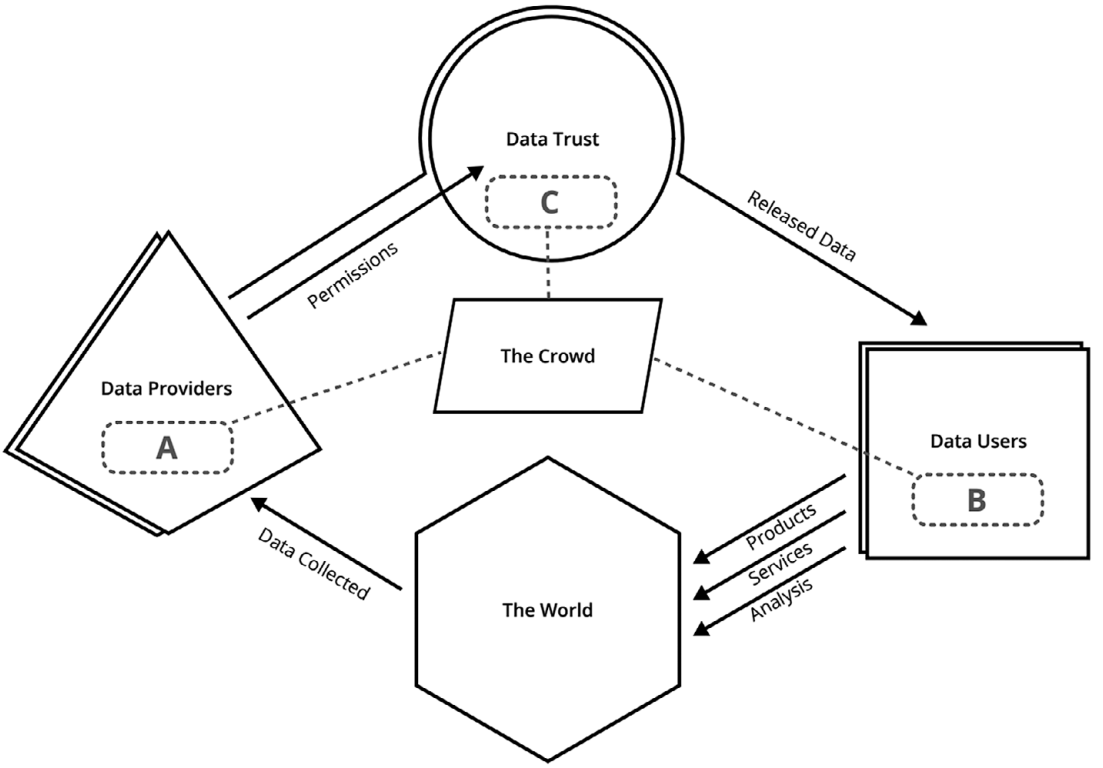

Figure 1. *Crowds within the data trust ecosystem, based on Open Data Institute (2019a).*

[Delacroix and Lawrence, 2018]), but they are not necessarily. In practice, it may even be rare that a single individual is both a data subject (of a dataset) and the data provider (to a data trust); and of course, not all datasets contain personal data, and hence many don't have data subjects at all.

The Crowd may be engaged in relationships with either the data provider or data user. For instance, crowd workers may be employed by a data provider during the creation of a dataset, prior to entrustment ("A" on the diagram). Or, they may be employed by a Data User as part of the process of deriving value from a dataset that is controlled by a data trust ("B" on the diagram); an example might be labeling a dataset in order to develop a machine learning model.

Typically, crowd workers are employed through an intermediary platform like Amazon Mechanical Turk which has a long-term relationship with both workers and employers, and which facilitates a short-term relationship between employer and worker through an online market (Howcroft and Bergvall-Kåreborn, 2019). The crowd platform provides infrastructure such as worker registration, worker reputation systems, and payment processing.

Finally, there is the possibility of a relationship between The Crowd and a Data Trust ("C" on the diagram). For the crowd, a governance stake might be based on the work that the crowd had contributed to the creation of the dataset.

We draw on Stakeholder Theory, and in particular the classification of stakeholder theories provided by Miles (Miles, 2017). Miles acknowledges that stakeholder theories are diverse and contested—that there is no single or agreed upon definition of what constitutes a stakeholder—and describes four primary categories of stakeholder theory, which describe different theoretical approaches to defining what (and hence, who) stakeholders are: Collaborator, Recipient, Influencer, and Claimant.

Influencer definitions "require a capacity to influence the operations of an organization and an active strategy to do so." Of all the stakeholder groups those that fall into Influencer definitions have the greatest power and interest in influencing a firm. Crowd workers do not fall within the influencer definitions.

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

Collaborator definitions "identify stakeholders through their ability to co-operate with organizations, regardless of their power to influence or the existence of a claim on corporate resources," but a collaborator stakeholder "lacks an active interest to influence the organization." Crowd workers can't be said to meet this definition of a stakeholder either, as they have a significant interest in influencing the platforms and requesters that they work with to—for instance—attain better pay or conditions. It's notable, though, that crowd work platforms often attempt to frame workers as being self-employed (Howcroft and Bergvall-Kåreborn, 2019), and thus might be regarded as attempting to frame crowd workers as collaborator stakeholders.

Claimant definitions "require a form of entitlement," (a claim). Claimant stakeholders actively pursue their claim, but "lack the coercive power to guarantee that their claims are attended to as their claims are derived from moral, social or residual rights rather than legal, direct, or economic claims associated with the power to influence." Crowd workers potentially fall into this category. They can, and do, communicate with platforms and requesters, but don't commonly have any means of forcing platforms or requesters to respond to their requests, or even to enforce compliance with specific contracts or laws.

Recipient definitions recognize asymmetrical relationships, to include any groups that are affected by a firm's actions. Mills explains that "these definitions imply no claim and no ability to influence." The agreements between crowd workers, platforms, and requesters give crowd workers a claim (at least theoretically) against those bodies, elevating their status beyond mere recipients; but a Data Trust could recognize its ability to affect crowd workers (as recipient stakeholders) and, for instance, require that data users pay a minimum wage to crowd data workers who are employed in processing a released dataset.

Data providers, data trusts, and data users could be subject to claims from various stakeholder groups, that are propagated through the legal agreements that bind parties together. Claims could arise from statute (e.g., the rights of data subjects), from contractual limitations imposed by one of the parties (e.g., a data provider limiting what purposes the deposited data may be used for), or commitments made to other parties—for instance by a data provider to crowd workers or citizen scientists—that are involved in the data value chain. In fact, any actor within the data value chain could be required to attend (in some form) to the claims of other stakeholders that have accrued throughout the chain.

In this research, one of our aims is to understand the possible claims or responsibilities that a Data Trust might need, or choose, to recognize in respect of crowd workers, and the possible implications of those claims on the structure and operation of data trusts.

3. Data Trust and Data Co-Op Survey

We were interested to see which stakeholders are currently represented in contemporary discourse and praxis around novel data governance models, and in particular, whether crowd data workers are among them. We undertook a survey of existing and proposed data trusts and data co-ops, using Google Web search and academic publications, respectively.

3.1. Methodology

Our methodology had four steps: identification of relevant examples and literature, following up on identified leads to obtain more information, applying inclusion criteria to narrow down the field of included examples, and summarization of the remaining proposals and examples. For brevity, we refer to the identified proposals, organizations, pilots, and products simply as "projects."

Initial identification: We conducted a survey of data co-ops using Google Web Search. The organizations that we identified are either self-described or described by others as "data co-ops" (or similar terms, such as "data co-operative"). Although relying on Google search results lacks reproducibility, in the absence of any other directory of data co-ops or trusts, it is a pragmatic way to identify organizations of interest. The full list of search terms used was as follows: "data co-op," "data co op," "data cooperative," "data co-operative," and "data trust." We supplemented our Web search results with academic literature that describes data trusts and co-ops using the same keywords in Google Scholar.

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

Following up leads: We found that many references to the keywords of interest were via secondary sources, such as news or trade websites, and we endeavored to follow those leads up either by following the provided links (where possible) or by searching for the names of projects that were mentioned.

Applying inclusion criteria: We have included projects in our survey where they meet (or, in the case of proposals, would meet) the following conditions:

1. the organization governs the use of a data resource of some kind; and
2. more than one party has (or can obtain) access to the data in the normal course of the project's operation.

Summarization: We created a summary of key information about each identified project, aiming to identify key aspects based on factors that are referred to in the existing literature about data trusts and co-ops. For each organization we aimed to identify:

1. The data domain or market sector;
2. the contents of the data store;
3. the aims and objectives or creating the data trust/co-op;
4. membership of the organization;
5. non-member beneficiaries; and
6. current status: For example, proposed, operational, or defunct.

A summary of all the trusts and co-ops that we surveyed is provided in Appendix 1.

3.2. Results

We discovered a mixture of proposed co-op/trust models and existing ones, and a great deal of diversity in aims, objectives, and domains. It is notable that many of the organizations we identified, which are called co-ops or trusts, do not meet the definitions of worker co-operatives or data trusts as described in the literature. The newness and the flexibility of terms and concepts are, still, a feature of the Data Trust design space.

We found that, of the projects surveyed, many did not involve any stakeholders that could be described as "workers" as members or beneficiaries. Instead, those projects pool data from multiple organizations in a "B2B" (business-to-business) relationship; making the term co-op somewhat broader in relation to data than it is generally understood in the context of labor.

Drivers' Seat is an example of a co-op that does include workers as members, in this case, gig-economy drivers who contribute data about their rides to help predict demand. Other projects (Savvy CoOp, Health Data Co-operative Model [HDCM]) include data subjects who might also be considered in some sense analogous to the workers in a labor co-op.

In some cases, the surveyed projects have an external beneficiary, that is, a stakeholder group that benefits (or is anticipated to benefit) despite not being involved in the running of the project. Non-member beneficiaries are a feature of legal trusts, but our results do not show that the existence of an external beneficiary is a differentiator between projects that are described as data trusts and projects that are described as data co-ops. For instance, the Ontario Forest Data Co-Operative has a clear public benefit, as does the HDCM; and the three ODI pilots also identify potential benefits to non-member stakeholders.

In answer to our main research question, of whether extant trusts and co-ops identify crowd workers or data workers as a member or beneficiary, the answer is no—none of the surveyed projects included crowd data workers as members or beneficiaries. Our results show that the inclusion of crowd data workers as identified stakeholders in a data trust would be a novel departure from current practice.

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

4. Crowd Worker Questionnaires

We are mindful of the need to include the voices of crowd workers in the design of future data governance models that will shape their work directly, or govern the exploitation of the results of their work. We wanted to see how crowd workers felt about the trade-offs that are inherent in different ways of governing and exploiting datasets, and what factors they feel are important in judging which models are preferable to others.

4.1. Methodology

Using data trusts as a structure for controlling the use of crowd-produced datasets has potentially far-reaching implications and possibilities for the form and structure of crowd work. For instance, if the capital required to produce the dataset is supplied in the form of crowd labor rather than as financial investment from a funder, the reward for a crowd worker is potentially uncertain and certainly time-delayed. Moreover, if deployed in a "co-operative" type model crowd workers themselves might be placed in a position to make decisions about how the data set is used.

We wanted to understand (a) how crowd workers would react to the sort of reward model that might be possible using data trusts (profit-sharing or royalties, but delayed and/or uncertain), and (b) how crowd workers might make decisions about the terms under which a crowd-produced dataset could be used.

To answer those questions, we conducted two online surveys. Participants in both surveys were crowd workers, recruited using Figure Eight, who were paid $1.50 for their participation upon completion of the task.

In both surveys, we collected some basic demographic information including age, gender, country of residence, time spent doing crowd work, income from crowd work, and the importance of crowd work income to the respondent. A full list of questions is provided in the supplemental materials that accompany this paper.

For both surveys, our analysis was primarily qualitative and took the form of a thematic analysis of the free-text questions. The design of the remainder of each survey is described next.

4.1.1. Survey 1: worker perspective

Our first line of inquiry was to consider how *fair* workers would consider a range of different scenarios to be. The scenarios were designed to include the implications of using a data trust to control the exploitation of the produced dataset. In this survey, we presented scenarios from the perspective of crowd workers.

We chose to frame our enquiry in terms of fairness because it is a widely understood concept that provides a simple way to elicit relevant opinions from crowd workers, without needing to explain potentially unfamiliar concepts such as stake or meaningfulness. As a relatively broad concept, it also invited participants to explore aspects of crowd work beyond just remuneration.

We presented three scenarios covering the exploitation of a crowd-classified medical imaging set, useful for developing a new diagnostic tool. The scenarios correspond to three different possible ways of licensing the resulting dataset, with different implications for the crowd workers.

Scenario A (Proprietary data): Our "baseline" scenario, which represents the status quo; a private company will exploit the results for private gain, workers receive an immediate fixed payment.

Scenario B (Royalties): This scenario represents a simplified data trust, in which the dataset is exploited for private gain; workers receive a small initial payment plus a share of the profits. Total (5 years) remuneration is higher than in Scenario A. Although profit-sharing is only one model of remuneration that's possible for a data trust to adopt, in this survey we were primarily interested in exploring the notion of delayed payment to workers.

Scenario C (Open Data): This scenario represents public-sector exploitation plus publication of Open Data. Workers receive an immediate payment (equal to Scenario A) but results are exploited by the public sector and made available to hospitals "at cost," and the dataset is published freely for others to use if they wish. In this scenario, we explore notions of public benefit.

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

We asked participants to reflect on the fairness of each scenario, to indicate what they felt was fair or unfair about each one, and to indicate and justify which scenarios they preferred. By framing the exercise comparatively and including three distinct scenarios we avoid the acquiescence bias that could have colored responses if we'd just asked about a data trust-type model.

4.1.2. Survey 2: trustee perceptions

We also wanted to understand how crowd workers might choose to exercise the rights that they'd vested in a data trust. Whereas Survey 1 placed participants in the (familiar) position of being a crowd worker, Survey 2 shifted the focus of the questions and scenarios to take the perspective of a "data trustee." We asked participants to consider how a dataset *should* be used by a series of different organizations. Allowing crowd workers a say in how datasets are governed is a viable proposition, but in this experiment, the framing is primarily of interest as a means of eliciting responses from our participants. We do not mean to suggest that crowd workers will, necessarily, be required to make this kind of decision in practice.

We asked participants to read a description of a (fictional) dataset—a set of labeled medical images that had been produced by crowd workers. We then presented participants with short profiles of four different (again fictional) organizations and asked participants to indicate whether they thought each organization should be allowed to use the labeled dataset, and how much they thought the organization should pay to use it. We wanted to capture any other factors that crowd workers would take into account when making their decision, so we asked if there were any questions that they'd like to ask the organization to inform their decision. Finally, we asked them to justify the decisions that they had made.

In this survey, we were particularly keen that participants' answers would be internally consistent, and so the profiles and questions about each organization were presented at the same time; participants could easily scroll up and down the page to refer to previous answers. To counteract ordering effects, the four sets of organization profiles and questions were presented to each participant in a random order.

The organizations were as follows (the full description provided to participants is included in the supplemental materials).

MediRay: A medical imaging company, based in the USA, that will use the dataset to develop a new diagnostic tool to sell to hospitals. A total budget of $8m, for an expected five-year profit of $25m.

BioViz: Similar to MediRay, but based in Sweden, BioViz will donate five of their developed diagnostic machines to hospitals in low-income countries (worth $2m). A total budget of $8m, for an expected profit of $20m.

Diagnostico Unido: Similar to MediRay, but based in Brazil. A budget of $5m for an expected five-year profit of $18m. We expected that a majority of participants would be from South America and wanted to see whether they would respond favorably to a more local company.

Rare Disease Research Lab: A publicly funded university research lab that will use the dataset to research the rare disease, but not to create new medicines or technologies directly; their results will be shared with other scientists freely. A budget of $500,000 for no expected profit.

4.2. Results

We begin by summarizing the participants in the two surveys, and then present the qualitative results of the two surveys in turn.

In total, 68 participants completed the surveys (32 and 36 in surveys 1 and 2, respectively). In survey 1, 13 were female and 19 were male; in survey 2, 13 were female and 23 male. The mean age in survey 1 was 36.2 and in survey 2, 31.6. In both surveys, the largest single country of residence was Venezuela, with a long-tail of other countries. Participant characteristics are shown in Table 1.

In both surveys, participants covered a range of household incomes and the income from crowd work varied from being a necessity to being just a nice addition. Our participants cover a range of backgrounds and experiences which makes them suitable for a qualitative exploratory study. Across the two surveys, participant demographics are fairly comparable, although survey 1 included several participants with household incomes above USD$50,000 per year, which survey 2 did not.

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

Table 1. Participant summary data

	Survey 1	Survey 2
Total participants	32	36
Age	36.2 (SD = 11.7)	31.6 (SD = 8.46)
Gender—female/male	13/19	13/23
Hours per day	Mean = 7.6 (SD = 3.8)	Mean = 6.0 (SD = 3.7)

Which statement best describes how important money from crowd work is to you?

	… always necessary to make basic ends meet	… sometimes necessary to make basic ends meet	... a way for me to pay for nice extras	... nice, but doesn't really change my circumstances
Survey 1	10	6	9	7
Survey 2	10	8	16	2

For how many years have you been doing crowd work regularly (at least three times per month)? (Years)

	<1	~1	~2	~3	~4	~5+	Non-regular
Survey 1	6	4	8	5	2	5	2
Survey 2	11	5	8	11	0	1	0

What is your annual household income? ('000 USD)

	<10	10–20	20–30	30–40	40–50	50–60	60–70	>=70
Survey 1	19	3	1	2	0	3	2	2
Survey 2	25	5	1	2	3	0	0	0

Although 68 participants is not a small sample for a qualitative study, the number of participants in any particular demographic category is quite small and so we have treated the results from each survey in aggregate and not considered how results differ between sub-groups.

During the qualitative analysis, we found that in general there were not clear differences between the themes that participants mentioned in response to different questions, and that it was, therefore, most instructive to consider the answers as a single corpus. Where themes do have a strong relation to a specific question or scenario, we have noted it.

4.2.1. Survey 1

We identified five main themes in the responses to survey 1, which we describe below. We have included illustrative quotes for each theme.

Theme: pay. Workers very frequently mentioned the amount that they would be paid for completing the task. Pay was typically compared either to the typical rate workers receive—for example, "Few jobs pay $8.50/hour, so I think this is a good scenario," (P20)—or to the rate of pay available through other jobs—"the pay is much better than what is gained in my country per hour," (P24) "the hourly pay is very good, considering my current income level," (P27).

Reflections on overall pay were primarily positive, reflecting the values offered in the scenarios. Although the amount of pay was perceived fairly positively across the scenarios, participants were somewhat

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

skeptical of delayed payments (i.e., the profit-sharing model in Scenario B). There seem to be two reasons for skepticism, which—at the least in the responses we received—were hard to distinguish: *Risk* and *Discounting.*

Theme: risk

By risk, we mean the inherent uncertainty in a deferred-payment model. One participant summed it up as

> "The payment sounds fair, the profit distribution I would not count on" (P21, of Scenario B).

> "I don't think the workers [want to] take the risks. They want sure money." (P12)

Risk seems to stem both from the risk inherent in an investment:

> "the future is uncertain" (P2, of Scenario B)

and from skepticism over the process for distributing future profits:

> "If any of workers loose their account (deactivation by some reason) they didn't receive money even if they completed [a] good job." (P3, of Scenario B).

Some participants commented on the level of pay if the risk didn't pay off:

> "It is risky. 1.50 per hour could be called slavery" (P19, of Scenario B).

Theme: discounting of deferred rewards

We saw evidence that some participants simply prefer an immediate reward, even over a higher deferred reward, which we refer to as "discounting" in line with economic descriptions of the same effect:

> "I would be interested in the immediate higher payment, so I can choose how to use my earnings" (P21),

> "The initial payment is very little, you would have to work several years to pay well" (P32).

Some participants preferred a higher but deferred reward, though;

> "...because in this I can earn more money in the long term if everything goes well" (P31)

Notably, P31 had indicated that income from crowd work was necessary to make ends meet.

Theme: distribution

As we might predict from prior work on fairness, participants felt that the payoff for the requester relative to the workers, that is, the *distribution*, was also an important aspect of the scenarios.

> "The company is going to earn a lot of money, and workers very little," (P32, of scenario A).

> "The payment $8.50/hour is fair for this type of job, no matter the profit of the company," (P21, of Scenario A).

> "It is fair for the company to distribute profit to the very men that helped" (P6, of Scenario B).

Some comments alluded to the legitimacy of a difference in earnings between worker and requester

> "because the company will earn a lot of money, while the workers will not, but even so it is a bit fair because the biggest gain should be from the company." (P31, of Scenario A).

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

"It could be thought that the pay is low considering the projected level of income. But objectively analyzing the matter is only part of the cost," (P27, of Scenario A).

One participant considered whether Scenario C was actually distributively unfair to the university involved:

"University does not profit. Maybe if it makes a little profit." (P22, on what's unfair about Scenario C).

Theme: public benefit

Beyond the distribution between worker and requester is the benefit to the public at large. Not surprisingly, public benefit was seen as a positive aspect of Scenario C. The idea of public benefit was often connected with worker motivation.

"The one that I liked the most so far. Fair salary for collaborators and a social action … I would feel better doing something that I know helps people." (P19).

"non profit work is always positive and I am willing to support it and share with reasonable pay" (P26).

"I think what the education is the basement of everything and we have to support all the research that they propose." (P20).

"The purpose of work allows the employee additional motivation" (P27).

4.2.2. Survey 2

Our analysis of Survey 2 was also primarily qualitative, but we also wanted to consider how much participants indicated they would charge each fictional organization to use the dataset. There was a large variation in the suggested amounts, and so rather than compare the absolute figures suggested, we ranked the amounts suggested for each organization and then compiled the overall rankings for each organization. The results of that ranking are shown in Figure 2.

To test the significance of the ranking differences we carried out a Friedman test comparing the payment rank for the four licensees to yield a highly significant Q value of 30.61, $p < .0001$. Six pairwise comparisons were then conducted using two-tailed Wilcoxon signed-rank tests. All pairwise comparisons (Table 2) are significant, with the exception of the comparison between Diagnostico Unido and BioViz, and remain so after a Bonferroni correction for multiple tests.

Figure 2 shows that RDRL (the university research lab) was usually ranked lowest, that is, most of our participants wanted to charge RDRL the least. Diagnostic Unido ranked third but only narrowly below BioViz, and MediRay was usually ranked highest. We consider the reasons behind the rankings in the discussion section.

Our thematic analysis of Survey 2 identified four main themes, which are similar to those in Survey 1.

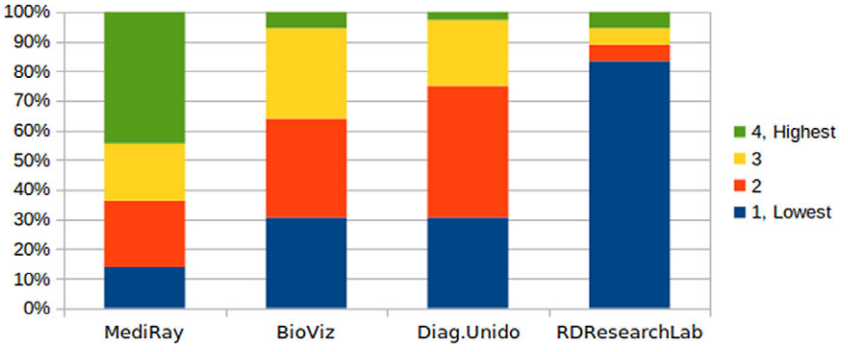

Figure 2. *Summary of how organizations ranked when ordered based on the charges.*

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

Table 2. Pairwise comparisons of payment ranks for Survey 2 scenarios

	BioViz	RDRL	MediRay
Diag. Unido	$T = 115.5$ ($p = .5$)	$T = 81.0$ ($p = .009$*)	$T = 23.0$ ($p < .001$*)
Mediray	$T = 46.5$ ($p = .001$*)	$T = 15.0$ ($p < .001$*)	
RDRL	$T = 84.5$ ($p = .007$*)		

Theme: pay-to-profit
Most participants felt that the organizations who were profiting should pay more to use the data.

> "The profit that the company will get is very big, so I think that this price is fair," (P246, of MediRay).

Conversely, RDRL, should get access to the dataset more cheaply,

> "because they will not get any kind of economic benefit and are in favor of health" (P244, of RDRL), and

> "it will not profit for the company, besides its image being improved in a world scenario, I think the price to be paid [a low amount] is fair" (P246, of RDRL).

This theme is consistent with the ranking analysis, and is similar to the distributional concerns in Survey 1.

Theme: public benefit
As in Survey 1, participants in Survey 2 raised public benefit as a positive aspect and one that cause them to reduce their proposed charge. For instance:

> "with my work both patients and the company would benefit… Any research for the benefit of humanity is useful." (P234, of all scenarios).

> "They are not only paying decent amount to workers, but are also during [sic] charity. Win Win situation." (P299, of BioViz).

> "As they are going to donate some of their time and effort and also are going to donate some equipment, I think we (the crowd workers) should reduce of incomes as well in order to support their effort." (P212, of BioViz).

Again, the idea that public benefit lowers the charge that participants suggested seems to be borne out by the ranking results where BioViz ranked lower than MediRay.

Theme: reward time and effort
As well as considering if the organization would profit, participants also considered whether the payment would fairly compensate workers for their time and effort.

> "Almost half the budget they have, a lot of hard work and time was spent to create the dataset and it should be rewarded properly." (P217, on Diag. Unido).

Unlike the "pay-to-profit" theme, which generally worked in favor of RDRL, participants had reservations over whether RDRL's small budget was sufficient to adequately compensate workers:

> "Crowd workers have worked hard for this data and should be rightly paid for their work" (P299, of RDRL).

> "I'm not sure the company could afford to pay workers well" (P277, of RDRL).

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

Theme: geographic impact

Several participants raised questions about the geographic impact of BioViz's promised donations, which we interpret as interest in how the public benefit will be distributed geographically. One participant said that they'd like to

> "request more information from the health centers that will receive the donated machines; economic conditions, location, etc." (P274, of BioViz).

Another raised several questions: related to geographic impact, although encompassing commercial reach as well as a public benefit:

> "Do you intend to sell the machine to just hospital in Brazil or worldwide? Is the price of the machine for low or mid-income countries lower than US or rich countries? If the profit you got from other countries excesses the expected profit, will you donate more machine for other lower-income countries?" (P233).

5. Discussion

The results from the questionnaires are broadly supportive of one another, and there are very clear similarities between the themes that we identified in them.

We find that, as prior work on fairness would suggest, the overall *distribution* of reward from the data work activity is important to crowd workers. Our participants expected organizations that intended to profit from their use of the resultant dataset to pay more, and clearly used the amount of expected profit in their justifications of how reward should be allocated.

Our participants were prepared to lower their charges for organizations that intended to deliver *public benefit*, either in part (BioViz) or as their sole aim (RDRL). They also suggest that knowing about a public benefit in advance would give additional motivation—or meaning—to their work.

However, the global nature of crowd work has some implications for public benefit concerns. Participants raised the *geographic distribution* of the impact arising from dataset use—including public benefit—as something they would like to consider. It may be more fruitful to think of the beneficiaries of crowd-produced datasets as plural "publics," and to organize data trusts in a way that acknowledges crowd workers' desire to see the benefits distributed around the world.

The relatively lower charge expected from Diagnostico Unido (DU) may be evidence of our participants considering geographic impact and weighing it favorably, especially in light of the relatively large number of participants from South America. However, the DU scenario also included a lower expected profit, and so the effect could be solely due to the "pay-to-profit" effect.

The results of Survey 2 also suggest that crowd workers expect to be *remunerated fairly in return for their time and effort*, even where their work contributes to a public benefit; and we suggest that this factor probably sets a lower bound on what they are prepared to accept in exchange for granting access to a crowd-produced dataset.

From our empirical results, we reflect on the affordances of datasets for granting long-term stake, point to the organizing principles of data trusts as an opportunity to change the interactions between stakeholders and disrupt extant models of crowd data work, and identify open design challenges for Data Trust creators.

5.1. Datasets as Pragmatic Vehicles for Crowd Stake

Datasets are potentially valuable capital resources, in a wide range of domains; and the results of the survey of existing data co-ops demonstrate that multi-stakeholder constellations can coalesce around data assets for mutual gain, for commercial or non-commercial reasons. Data's status as an asset, subject to property rights, differentiates crowd data work from other forms of labor in the "gig economy." Whereas the outputs from service-based crowd work like food delivery are intangible and ethereal, the asset created

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

by crowd data work—the resulting dataset—can be accounted for, enriched with provenance metadata, and made subject to legal restrictions that protect the interests of stakeholders.

Data trusts' status as bodies that can influence the use of data assets gives them, quite pragmatically, an ability (whether they choose to use it, or not) to influence the dynamics of the crowd data work. Crowd data workers are, at least, recipient stakeholders of data trusts; and it is not difficult to imagine that the legal agreements between workers, employers, and data trusts could create enforceable legal claims for workers, that would elevate them to Claimant stakeholder status.

5.2. New Logics and Arrangements

Our exploration of data trusts from the point of view of crowd workers also gives us pause to reflect on the "organizing principles" of data trusts; their fundamental guiding logic. The status quo of crowd microwork frames workers as "collaborator" stakeholders, who voluntarily engage in work in return for a wage. The organizing principle is capitalistic in nature; the reward primarily accumulates to the data owner, often creating poor and unfair conditions for workers. However, those arrangements embody what Bowles and Gintis call the "contradictory nature of liberal democratic capitalism," and bring into contrast the organizing principle of capitalism—accumulation deriving from property ownership—with the organizing philosophy of liberal democracies which "vests rights in persons" (Bowles and Gintis, 1978). Data trusts disrupt this model through their pursuit of social value, and a focus on trustworthiness and ethics; O'Hara terms it an "ethical regime" and the ODI describes as a "responsibility to take the interests of data holders, data users, citizens, and other stakeholders into account" (Hardinges et al., 2019).

The organizing logic of a trust, which represents the interests of others rather than of itself, is potentially disruptive because the data trust may become a locus for the dispersed stake of many crowd workers to be consolidated; there's no inherent tension between the Trust itself—guided by its ethical regime more than commercial self-interest—and the stakeholders that it represents; albeit that there could be tensions between stakeholders that the Trust must navigate, or even arbitrate.

Our results show quite conclusively that crowd microworkers are positive about being involved in data work shaped by that new organizing logic; that they view their work as more than just a source of income and consider social impact and equity as positive factors to be weighed against their own remuneration.

The shift in organizing principles, from the commercial return in the status quo to the broader social outcomes of a Data Trust, opens the possibility of new data governance "patterns" that better suit the broader objectives; and those patterns have direct implications for crowd data workers.

The simplest new arrangement is for a data provider to employ crowd workers per current practice ("A" in Figure 1), but then to grant the dataset to a data trust at some later point. The main benefit to crowd workers in this arrangement is closely aligned to the Trust's own goal of making greater use of the dataset than might be possible if it were only available to the provider; that social outcomes valued by the crowd workers are increased.

As we speculated earlier, there's also the possibility of more direct relationship between crowd workers and the data trust ("C" in Figure 1). Such a relationship could be part of giving value back to crowd workers directly, for instance through a delayed-remuneration scheme as we described in our scenarios. That relationship might exist where a dataset is provided to the trust by a data provider (as above), but feasibly the data trust could itself take on aspects of a worker co-operative. Such a model would be compatible with the organizing logic of a data trust, but (as we show in our results) raises some challenges around worker trust and risk that might be partially solved through design interventions.

The relationship between data trusts and data users gives data trusts a lever through which to control "downstream" employment of crowd workers (shown as "B" in Figure 1). As part of their ethical regime, data trusts could design data licenses or other legal mechanisms to hold data users to standards around crowd data worker employment. Data trusts/co-ops as reactive, "living" governance mechanisms—compared to "publish and forget" data sharing models—can balance rights contextually, similarly to how O'Hara (2019) argues a data trust might audit and hold to account data scientists who are granted access.

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

5.3. Design Process Challenges

Throughout our research, we've deliberately framed the creation of data trusts as a design challenge, recognizing that the process of creating a successful trust will require careful consideration of different stakeholder needs and the creation of a variety of tangible and intangible artifacts that enable the trust to realize its intended outcomes. Engaging with crowd workers poses some particular challenges for a design process, though.

Crowd workers are not a homogeneous group; there is a great deal of geographic and demographic diversity, as shown in our own results. Furthermore, the composition of the crowd is not stable over time. Previous research by Ross et al (Ross et al., 2010) noted a shift from US-based workers to those in India, and the participants in our research were primarily from Venezuela. It will be important for data trusts to identify common features, but also to recognize the diversity of the crowd and respond to changes in the needs and wishes of the crowd workers. Crowd worker expectations for the geographic distribution of social benefits are one example of a factor that may vary between different crowds.

The crowd is also diffuse; whereas an employer might consult with unions, and form working groups from its pool of known employees, there aren't obvious representatives or bodies that can act as a sounding board on behalf of the crowd.

The solution to both these challenges may lie in creating dynamic platforms and organizations that include mechanisms for crowd workers to shape key decisions. For example, the Daemo project (Gaikwad et al., 2017) scaffolded a general iterative process that allowed crowd workers to improve the quality of tasks based on factors important to the crowd workers themselves. By extension, such a method is able to change in response to the shifting needs of the workers themselves. Similar consultative mechanisms could be applied to shape licensing terms, to contribute to shaping the trust's ethical regime, or to set the purposes to which a dataset may be put.

5.4. Design Challenges

Our results, in the context of the new logic and arrangements made possible by data trusts, also suggest some design challenges for the creators of novel data governance models; challenges that might be overcome through a combination of governance design, incentive design, and tool design.

5.4.1. Risk

Risk was a cross-cutting concern for our participants. Risk and uncertainty are inherent in a data trust model, where data exploitation and possibly worker remuneration take place in the future and may not have been fully defined at the point that the dataset is created.

Our participants identified risk as stemming from two sources: uncertainty about future exploitation (and hence profit/benefit)—a typical capital investment risk—and procedural uncertainty that stems from the time delay and administrative chain between doing the work and being rewarded.

There was disagreement among our participants about their appetite for risk. Some participants felt negative about taking a risk on the profit-sharing model proposed in Survey 1, while others reacted positively to the idea of receiving greater rewards in the future as a result of it.

The degree of risk that workers are exposed to is likely to vary based on how a specific data trust operates. In Survey 2, the proposed model (a licensing fee, rather than profit-linked royalties) exposed workers to less inherent risk and we found that our participants did not frame their answers in terms of risk or uncertainty. Whereas in Survey 1, the profit-sharing model places a similar risk on workers as it would on providers of financial capital.

Requesters that intend to place crowd outputs into a data trust will need to find ways of conveying risks around uncertain payoff to crowd workers to enable them to make informed decisions about the projects they contribute to. Conveying risk should not just emphasize the legitimate risks, but also reassure crowd workers about potentially unfamiliar arrangements that—as our results show—are perceived as inherently riskier.

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

5.4.2 Managing long-term stake

Risk that stems from a time delay may be unavoidable in a data trust (or co-op) model, where workers take a longer-term stake in the dataset that's produced. Data trusts will need to be able to reassure workers that their long-term commitments can be honored, and there are perhaps opportunities to explore how concepts such as "digital wallets" (a concept now embedded in digital currencies) or other persistent digital identifiers can make long-term stake-holding possible across jurisdictions, and in countries where access to formal ID may be relatively difficult.

5.4.3 Peer-to-peer trust

Uncertainty about future exploitation is compounded by pooling many workers' contributions together and rewarding them based on aggregate success (e.g., demand for the overall dataset, or profit derived from it) rather than (as is currently the case) rewarding them based on their own individual contribution.

One participant summed up their feelings about being reliant on other crowd workers by writing "I would not work in B, knowing that I have to depend on dishonest workers to collect a 'living wage'" (P19).

Just as crowd work platforms have developed sophisticated means to score workers based on experience and past performance for the benefit of employers, if crowd workers are to contribute their efforts to a dataset held in trust, they may need to know something about their peers. Exploring what would reassure workers about contributing their efforts to such a dataset could be a fruitful area for further research. Examples such as eBay show how reputation systems can build trust among strangers, although building one-to-one trust between a buyer and seller is somewhat different from building trust between a crowd worker and the rest of the crowd.

6. Conclusions and Limitations

In this paper, we have described a survey of existing and proposed data co-ops and data trusts, and two participant surveys that explore crowd workers reactions to scenarios that could arise if the outputs of crowd data work—crowd-produced datasets—were held in data trusts that recognized crowd workers as stakeholders.

We have discussed how the organizing logic of a data trust differs from that of current purely capitalistic requester-worker relationships and how the new logic opens up the potential for more diverse models of crowd worker engagement.

The crowd's involvement in data work is, at the very least, a moral hazard that poses a challenge to the "ethical regime" of a data trust. Data trusts, able to consider and arbitrate the interests of a range of stakeholders, also have the potential to re-shape aspects of crowd work and to explore new models that support new data value chains and also improve conditions for crowd data workers.

Importantly, our results show that crowd workers recognize the stakes of other stakeholders within data value chains and also respond positively to the potential social value of their work, as well as the immediate financial compensation that they receive.

New models pose design challenges to creators of data trusts, partly because of the diffuse and heterogeneous nature of the crowd, but also because the lifespan of a data trust adds uncertainty around the realization of medium- and long-term outcomes; and because crowd workers may doubt the administrative processes around delayed rewards.

As our analysis shows, data governance models that respect the stake of data workers are not necessarily a simple replacement or addition to existing crowd work practices, but rather these novel models have implications for the way in which crowd data work is structured, presented to workers, and underpinned through technology and tools. We suggest several implications, in particular.

First, that novel models can facilitate a shift in how the stake of crowd workers is accounted for, but in doing so would require the introduction of longer term stake models that can maintain a link between worker and output for the duration of a dataset's life cycle. Longer term stake models are important for

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

financial remuneration, but also for demonstrating the social impact that stems from a dataset and facilitating crowd involvement in governance processes.

Third, that governance structures need to leave room for large-scale discourse and discussion; having potentially thousands of micro-contributors is quite different to representing the voices of a handful of manufacturers, for example, and so decision-making fora and mechanisms must be altered accordingly. The scale and characteristics of the crowd give us reason to investigate new ways of consultation, and "crowd-centered" design.

Our analysis contributes to a better understanding what crowd workers might want from the growth of novel data governance structures, how their stake can be situated within a broader discussion of how competing interests in data sets are balanced, and design and research implications for the technical underpinnings of these novel data governance models.

Funding Statement. This research was supported by funding from the European Union's Horizon 2020 research and innovation program under the QROWD grant agreement No 732194

Competing Interests. The authors declare no competing interests exist.

Data Availability Statement. The qualitative survey data that supports the findings is available at https://doi.org/10.5281/zenodo.3873333.

Author Contributions. Conceptualization, R.G., E.S.; Methodology, R.G., E.S.; Data curation, R.G.; Data visualization, R.G.; Writing an original draft, R.G.; Writing review & editing, E.S. All authors approved the final submitted draft.

Supplementary Materials. To view supplementary material for this article, please visit http://dx.doi.org/10.1017/dap.2020.21.

References

Borkin S (2019) *Solving the Capital Conundrum* (February).

Borooah V, **Glyn A**, **Miliband D**, **Goodman A and Webb S** (1995) Paying for inequality: the economic cost of social injustice. *The Economic Journal 105*(433), 1649–1651. https://doi.org/10.2307/2235125

Bowles S and Gintis H (1978) The crisis of liberal democratic: the case of the US. *Politics & Society 11*(1), 51–93.

Busarovs A (2013) Ethical aspects of crowdsourcing, or is it a modern form of exploitation. *International Journal of Economics & Business Administration I*(1), 3–14.

Cavanillas JM, **Curry E and Wahlster W** (2016). *New Horizons for a Data-Driven Economy* (Cavanillas JM, Curry E and Wahlster W (eds). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-21569-3

Craig B and Pencavel J (1995) Participation and productiviy: a comparison of worker cooperatives and conventional firms in the plywood industry. *Microeconomics* Vol. *1995*, pp. 121–174.

Dawson R and Bynghall S (2012) Build your business by tapping one of the most powerful trends in business today: crowdsourcing. In *Getting Results From Crowds.* AdvancedHuman Technologies Inc.

Delacroix S and Lawrence N (2018) Disturbing the "one size fits all," feudal approach to data governance: bottom-up data trusts. *SSRN Electronic Journal*, 1–30. https://doi.org/10.2139/ssrn.3265315

Ettlinger N (2016) The governance of crowdsourcing: rationalities of the new exploitation. *Environment and Planning A: Economy and Space 48*(11), 2162–2180. https://doi.org/10.1177/0308518X16656182

Fieseler C, **Bucher E and Hoffmann CP** (2017) Unfairness by design? The perceived fairness of digital labor on crowdworking platforms. *Journal of Business Ethics 156*(2), 1–19. https://doi.org/10.1007/s10551-017-3607-2

Fort K, **Adda G and Cohen KB** (2011) Amazon mechanical turk: gold mine or coalmine? *Computational Linguistics 37*(2), 413–420. https://doi.org/10.1162/COLI_a_00057

Gadiraju U, **Kawase R and Dietze S** (2014) A taxonomy of microtasks on the web. *Proceedings of the 25th ACM Conference on Hypertext and Social Media—HT'14*, pp. 218–223. https://doi.org/10.1145/2631775.2631819

Gaikwad Snehalkumar (Neil) S. Gaikwad, Mark E. Whiting, Dilrukshi Gamage, Catherine A. Mullings, Dinesh Majeti, Shirish Goyal, Aaron Gilbee, Nalin Chhibber, Adam Ginzberg, Angela Richmond-Fuller, Sekandar Matin, Vibhor Sehgal, Tejas Seshadri Sarma, Ahmed Nasser, Alipta Ballav, Jeff Regino, Sharon Zhou, Kamila Mananova, Preethi Srinivas, Karolina Ziulkoski, Dinesh Dhakal, Alexander Stolzoff, Senadhipathige S. Niranga, Mohamed Hashim Salih, Akshansh Sinha, Rajan Vaish, Michael S. BernsteinOrting: Silas Ørting, Andrew Doyle, Arno van Hilten, Matthias Hirth, Oana Inel, Christopher R. Madan, Panagiotis Mavridis, Helen Spiers, Veronika Cheplygina (2017) The Daemo crowdsourcing marketplace. *CSCW 2017—Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1–4. https://doi.org/10.1145/3022198.3023270

Gregg P, **Machin S and Manning A** (1994) High pay, low pay and labour market efficiency. In Glyn A and Miliband D (eds), *Paying for Inequality.* London: IPPR, pp. 100–113.

Hall W and Pesenti J (2017) *Growing the Artificial Intelligence Industry in the UK.*

Hardinges J (2018) What Is a Data Trust? Available at https://theodi.org/article/what-is-a-data-trust/ (accessed 1 May 2019).
Hardinges J (2020) Data Trusts in 2020. Available at https://theodi.org/article/data-trusts-in-2020/ (accessed 1 Oct 2020).
Hardinges J, **Wells P**, **Blandford A**, **Tennison J and Scott A** (2019) *Data Trusts: Lessons from Three Pilots. Open Data Institute*. London.
Homans GC (1961) *Social Behavior: Its Elementary Forms*. New York: Harcourt, Brace and World, Inc.
Hopkin A, **Fenech A**, **Liljalehto H**, **McLaughlin D and Williams T** (2001) The Ontario forest Health Data Co-Operative. *Environmental Monitoring and Assessment 67*(1–2), 131–139. https://doi.org/10.1023/A:1006474205223
House of Lords (2018) *AI in the UK : Ready, Willing and Able? Select Committee on AI: Report of Session 2017-19.* Available at https://doi.org/10.1016/j.jelectrocard.2010.12.041
Howcroft D and Bergvall-Kåreborn B (2019) A typology of crowdwork platforms. *Work, Employment and Society 33*(1), 21–38. https://doi.org/10.1177/0950017018760136
Kost D, **Fieseler C and Wong SI** (2018) Finding meaning in a hopeless place? The construction of meaningfulness in digital microwork. *Computers in Human Behavior 82*, 101–110. https://doi.org/10.1016/j.chb.2018.01.002
Miles S (2017) Stakeholder theory classification: a theoretical and empirical evaluation of definitions. *Journal of Business Ethics 142*(3), 437–459. https://doi.org/10.1007/s10551-015-2741-y
O'Hara K (2019) *Data Trusts : Ethics, Architecture and Governance for Trustworthy Data Stewardship* (February), 1–21.
Open Data Institute (2019a) *Data Trusts Summary Report.* Available at https://theodi.org/wp-content/uploads/2019/04/ODI-Data-Trusts-A4-Report-web-version.pdf (accessed 1 May 2020).
Open Data Institute (2019b) *Exploring the Potential for Data Trusts to Help Tackle the Illegal Wildlife Trade.*
Open Data Institute (2019c) *Exploring the Potential of Data Trusts in Reducing Food Waste.*
Open Data Institute (2019d) *Greater London Authority and Royal Borough of Greenwich Data Trust Pilot.*
Ørting S, **Doyle A**, **van Hilten A**, **Hirth M**, **Inel O**, **Madan CR**, … **Cheplygina V** (2019) *A Survey of Crowdsourcing in Medical Image Analysis.* Available at http://arxiv.org/abs/1902.09159.
Prassl J and Risak M (2016) Platforms as employers? Rethinking the legal analysis of crowdwork. *Comparative Labor Law & Policy Journal 37*(3), 619–652. https://doi.org/10.3868/s050-004-015-0003-8
Rawls J (1985) Justice as fairness: political not metaphysical. *Philosophy & Public Affairs 14*(3), 223–251. Available at http://www.jstor.org/stable/2265349
Ross J, **Zaldivar A**, **Irani L**, **Tomlinson B and Six Silberman M** (2010) Who Are the Turkers? Worker Demographics in Amazon Mechanical Turk. *Chi Ea 2010.* https://doi.org/10.1145/1753846.1753873
Sadowski J (2019) When data is capital: datafication, accumulation, and extraction. *Big Data and Society. 6*, 1, https://doi.org/10.1177/2053951718820549
Salehi N, **Irani LC**, **Bernstein MS**, **Alkhatib A**, **Ogbe E**, **Milland K and Clickhappier** (2015) We are dynamo: overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems, 18–23 April 2015, Seoul, Republic of Korea.* https://doi.org/10.1145/2702123.2702508.
Stalla-Bourdillon S, **Wintour A and Carmichael L** (2019) *Building Trust Through Data Foundations A Call for a Data Governance Model to Support Trustworthy Data Sharing*, 33.
Van Roessel I, **Reumann M and Brand A** (2018) Potentials and challenges of the Health Data Cooperative Model. *Public Health Genomics 20*(6), 321–331. https://doi.org/10.1159/000489994
Whiting ME, **Hugh G and Bernstein MS** (2019) *Fair Work: Crowd Work Minimum Wage with One Line of Code.* Available at https://github.com/StanfordHCI/fairwork.
Wortman Vaughan J (2018) Making better use of the crowd: how crowdsourcing can advance machine learning research. *Journal of Machine Learning Research 18*, 1–46. Available at http://jmlr.org/papers/volume18/17-234/17-234.pdf

Cite this article: Gomer R. C. and Simperl E. (2020). Trusts, co-ops, and crowd workers: Could we include crowd data workers as stakeholders in data trust design?. *Data & Policy*, 2: e20. doi:10.1017/dap.2020.21

Appendix 1

Summary of existing data trusts and data co-ops

Co-op/trust name	Data content	Aims/objectives	Membership	Current status	Ref/link
Proposals and archetypes					
HDCM	Health Data, submitted by subjects	Make data available for research, or to medical professionals for care/advice	"citizen owned and managed," that is, medical data subjects	Proposal	Van Roessel et al. (2018)
Cross-website advertising data sharing	Data about website visitors, for example, inferred demographics, market segments	Pool data between participants to support, for example, targeted advertising of site visitors	Participating website owners	Archetype	US Patent US10129323B1 (Google LLC)
Data Consortia Platform	Could contain any data, pooled by data subjects	A proposed archetype, which would see data subjects pool data to be licensed to potential data-consumers	Data subjects	Archetype, example given of "MIDATA" (Switzerland)	Borkin (2019)
Deployed examples					
Ontario Forest Health Data Co-Operative	Forest (tree) health data	Improve forest management and research	Canadian forestry organizations, including	Defunct	Hopkin et al. (2001)
Adobe Experience Cloud Device Co-Op	"device link information" that allows a user to be identified across multiple devices	To identify users across multiple devices, by pooling the "links" that different members are able to identify	Organizations that wish to share, and use, inferred "device links"	Deployed	https://docs.adobe.com/content/help/en/device-co-op/using/home.html
Driver's Seat	Data from ride-share drivers	Help drivers make informed decisions about their work and to generate income by licensing data to interested parties like local authorities	Ride-share drivers	Deployed, beta-testing	https://www.driversseat.co/
Savvy Co op	Data about patients; primarily contact details and health-condition information	Connect researchers/companies with patients who can provide insight into their conditions to, for example, inform product development	Patients, who get paid when they are invited to give their insights about health care to interested researchers/companies	Deployed	https://www.savvy.coop/

https://doi.org/10.1017/dap.2020.21 Published online by Cambridge University Press

Continued

Co-op/trust name	Data content	Aims/objectives	Membership	Current status	Ref/link
ADARA	Search, purchase and loyalty data about people who use travel services	Help travel companies to market their services and understand travelers	Travel companies; for example, airlines, hotel, car rental	Deployed	https://adara.com/
Bombora Data Co-Op	B2B "firmographics and Intent data"	Enable B2B sales & marketing by pooling data about firms and visitor consumption data	Organizations in B2B marketing supply and demand chain: for example, vendors, analysts, content syndicators, publishers, marketers	Deployed	https://bombora.com/data/bombora-data-co-op/
CoreLogic Data Co-Op	Property sales listings, plus data on neighbourhoods and so forth	Helps estate agents to access listing information from other areas to, for example, help clients who are relocating away, and to share their listings with other agents doing the same	Estate agents	Deployed	https://www.corelogic.com/products/data-co-op.aspx
GLA/Greenwich ODI Pilot	Data about electric vehicle parking spaces and data collected by heating sensors in residential housing	Aimed to improve data flows to help with two use cases: making use of electric vehicles more attractive by identifying available parking spaces, and improving energy-efficiency of social housing	Greater London Authority and Royal Borough of Greenwich	Pilot	Open Data Institute (2019d, 2019a)
WildLabs Tech Hub ODI Pilot	Data about the illegal international wildlife trade; for example, data acquired at borders and image/acoustic data	Increase access to relevant wildlife data in order to reduce illegal wildlife trade	Law enforcement, wildlife organizations	Pilot, early investigation	Open Data Institute (2019b, 2019a)
Food Waste ODI Pilot		Reduce food wastage by helping organizations to understand the full food supply chain	Retailers, food manufacturers, organizations working to reduce waste, researchers, policy-makers, regulators	Pilot, early investigation	Open Data Institute (2019c, 2019a)