WHEN LESS TESTING IS MORE

The Political Power of Experimentation in Data-Driven Organizations

Gretta Corporaal & Georg Rilinger

Abstract

In data-driven organizations, decision making is often treated as a neutral method for guiding product development, with leaders expected to resolve disagreements by appealing to "what the data shows." Yet in practice, data is rarely neutral, and coalition building often requires strategic ambiguity rather than clarity. This raises a puzzle: how do organizational actors forge coalitions with appeals to data when data is always contestable? Drawing on ethnographic fieldwork inside the platform organization iLabor, we show that project leads relied not on arguments about specific data, but on appeals to the *ideology* of experimentation: the belief that structured testing would ultimately yield objective insight. We find this ideology's authority is inherently unstable, and conceptualize this as epistemic fragility: a political resource that collapses under the very scrutiny it purports to invite. This ideology enabled three key strategies that helped project leads to build coalitions across groups with conflicting interests: asserting experimental authority, deferring conflicts, and promising future resolution. Our study reveals that the political power of experimentation hinges on a paradox: it is most effective when its actual practices remain peripheral and unexamined.

Introduction

Organizations increasingly embrace data-driven decision making. They adopt norms and processes to systematically collect and analyze data about their customers to produce, assess, and refine their products (Fourcade & Healy, 2024). Firms like Amazon, Facebook, and Google routinely conduct thousands of concurrent experiments to evaluate new features, test incentives, and fine-tune user interactions (Marres & Stark, 2020; Pinch, 1993; Rahman et al., 2023). This approach is often framed as a form of "scientific management," promising to replace managerial intuition with objective knowledge (Thomke, 2001).

Yet, organizations are not simply information processors; they are also political systems (March, 1962). Advancing projects requires the negotiation of compromises between groups

with divergent goals (Kellogg, 2011; Truelove & Kellogg, 2016). Research on data-driven decision making suggests it may help organizations to understand their market environment, but it also reshapes the political dynamics within them (Karp, 2025; Rerup & Zbaracki, 2021). Accordingly, leaders must mobilize data not only to identify the best course of action, but also to resolve conflicts and coordinate divergent interests (Huising, 2014; Kaplan, 2008). However, studies on quantification suggest that formal metrics lose authority precisely when they are drawn into political contestation (Christin, 2020; Mennicken & Espeland, 2019). In some cases, coalition building may even require strategic ambiguity rather than clarity to hide deep-seated misalignments of interests and secure buy-in (Leifer, 1991; Reinecke & Donaghey, 2025). This motivates us to ask: *"How do organizational actors build coalitions in data-driven organizations when arguments from data are inherently contestable and clarity is not always desirable? And under what conditions do these strategies fail?"*

We address these questions through a 15-month ethnographic study of coalitional politics at "iLabor" (a pseudonym), an organization operating one of the world's largest digital labor markets. Combining eight months of immersive onsite fieldwork with remote observation, we followed "Project MATCH," a strategic initiative to redesign the platform's algorithmic matching system. This structural reform required continued buy-in from groups with divergent interests: front-end personnel emphasized user experience, back-end engineers prioritized system integrity, and market management focused on aggregate outcomes. We analyze how project leads navigated the resulting goal conflicts.

Despite the organization's deep commitment to data-driven decision making, we find that project leads rarely resolved disputes by determining 'what the data shows.' Rather, they appealed to the *ideology* of experimentation: the shared belief that rigorous testing would ultimately yield

objective and decisive insight. This ideology enabled three strategies: "asserting experimental authority" to disqualify opposition, "deferring conflicts" pending evidence, and "promising future resolution" through experimental development. Crucially, these strategies proved effective only when actual experimental practices remained insulated from scrutiny. Early in the product cycle, experiments addressed marginal issues and produced seemingly self-evident results. But as the project advanced and experiments addressed politically central questions, the resulting data became ambiguous and practices came under scrutiny. This eroded the ideology's plausibility and project leads' ability to invoke it. We call this dynamic the *epistemic fragility of experimentation*: the conditional authority of experimental ideology collapses under the very scrutiny it purports to invite. As experiments move from periphery to center and results shift from clear to ambiguous, the gap between ideal and practice becomes visible, undermining the strategies that depend on experimental authority. Ironically, we thus find that political appeals to experimentation are most powerful when experimental practice is used sparingly - its authority depends on an idealized analogy between laboratory and corporation that is most potent when unexamined.

This study makes three key contributions. First, it extends research on the symbolic power of numbers by theorizing how the authority of quantification depends not on methodological rigor but on insulation from scrutiny. Second, we advance research on organizational politics by showing how appeals to experimental ideology can sidestep core problems of coalition building in data-driven organizations. Third, we contribute to research on data-driven organizations by identifying a political, rather than epistemic, reason for incrementalism: when data-driven decision making forces rather than defers political conflict, coalition building for structural change becomes more difficult.

Theoretical Background

Does Data-Driven Decision Making Displace Coalitional Politics?

Organizations are political systems (March, 1962). Members of different departments and subunits have divergent interests and therefore tend to pursue goals that conflict with those of other groups (Bechky, 2003; Cyert & March, 1963; Truelove & Kellogg, 2016). To advance projects, leaders must build coalitions among groups with different goals and interests (Koppman et al., 2022; Levinthal & Pham, 2024; Mithani & O'Brien, 2021). The literature on data-driven decision making typically treats this process as unproblematic, assuming that objective information tempers organizational politics.

Taking an information-processing view of organizations (Brynjolfsson et al., 2011; Camuffo et al., 2020), this literature views politics as the result of environmental uncertainty. With sufficient information, organizations will choose the best course of action. But when information is scarce, decisions are underdetermined and leave room for psychological bias and divergent subunit goals to assert themselves (Brynjolfsson & McElheran, 2019). From this perspective, if data collection and analysis improve access to reliable information, there simply is less room for political struggles (Kim et al., 2024; Sitkin, 1992; Thomke, 2020). When problems do arise, they are viewed as stemming from poor implementation (Luca & Edmondson, 2024; Ransbotham et al., 2016) or environmental myopia (Allen & McDonald, 2025; Felin et al., 2020; Wu et al., 2020).

This premise is problematic for two reasons, however. First, most studies model data-driven decision making at the organizational level: a monolithic organization runs an experiment or collects data and then decides whether to adopt a change (e.g. Koning et al., 2022). In practice, data

production and analysis occur within specific projects where members of different units collaborate (Hall & Hasan, 2020; Kohavi et al., 2009). Goal conflicts between these units are rarely just by-products of environmental uncertainty; they are endemic to the division of labor (Koppman et al., 2022). Departments occupy distinct structural positions with different career incentives and material interests, and they participate in separate cognitive and cultural communities (Dougherty, 1992; March & Simon, 1958; Nigam et al., 2016). Conflicts thus concern not only what is true but also whose perspectives and goals should take precedence. Second, when mobilized to resolve such conflicts, data seem not to function as an impersonal authority. Instead, each side tends to mobilize them as ammunition to advance their agenda. Data-driven decision making is unlikely to eradicate coalitional politics; rather, it appears to *reconfigure* them.

Research on coalition building has identified multiple mechanisms for settling goal conflicts (cf. DiBenigno, 2018). One line of work describes coalitions as outcomes of bargaining processes, in which groups trade support for inducements, or as temporary truces that are based on mechanisms such as sequential attention to goals, binding organizational structures such as budgets, slack, and other forms of administrative fiat (Cyert & March, 1963; Ethiraj & Levinthal, 2009; e.g. Greve, 2008). A second highlights integrative mechanisms that align goals through interaction and mutual adjustment, including cross-functional structures, integrator roles, and collaborative incentives (Levina & Vaast, 2005; e.g. March & Simon, 1958; Van Den Bulte & Moenaert, 1998). Both approaches require authority structures that organize negotiations, force compromises, and impose binding decisions in case of deadlocks.

Yet, in data-driven organizations, hard evidence is supposed to stand in for formal or occupational authority. When two subunits disagree, an analysis of data is meant to settle the debate. For example, subgroups able to demonstrate their ideas in an A/B test win arguments against

those who cannot (Kohavi et al., 2020). This shift fundamentally alters the nature of coalition building. Organizational actors must now persuade rather than decide; no longer able to impose outcomes through hierarchical fiat or negotiated side payments, they must win arguments with appeals to evidence. Viewed through the lens of the quantification literature, how this form of persuasion could be used to build coalitions remains a critical theoretical puzzle.

Unpacking the Politics of Numbers in Data-Driven Organizations

Numerical evidence carries a distinctive form of authority. It appears to depersonalize judgment (Porter, 1996) while appealing to deeply held cultural values such as objectivity, discipline, and fairness (Daston & Galison, 1992; M. S. Feldman & March, 1981; S. Feldman, 2004; Mennicken & Espeland, 2019). For instance, Carruthers and Espeland (1991) showed that double-entry bookkeeping gained influence not because it was more effective than alternative approaches, but because it was rhetorically appealing as neutral and objective. Organizations can therefore turn to the authority of numbers when trust in discretion is low (Bower, 1972).

Yet not all metrics are imbued with the cultural value of "mechanical objectivity" (Daston & Galison, 1992, p. 82). They gain authority by becoming embedded in organizational life, such as through infrastructures that organize comparison (Espeland & Stevens, 1998, 2008; Leonardi & Treem, 2020), alignment with shared meanings (Ranganathan & Benson, 2020) or ritualized production processes (Mazmanian & Beckman, 2018). Conversely, when metrics clash with professional judgment or lived experience, their constructed nature becomes visible, eroding their power (Anthony, 2021; Bechky, 2020; Levina & Orlikowski, 2009; O'Neill, 2016). For instance, experts may lose authority during 'censure episodes' that make their knowledge practices explicit and reveal mismatches with organizational goals (Huising, 2014).

While quantification scholars have examined how individual metrics gain authority (Mennicken & Espeland, 2019), less attention has been paid to how broader regimes of data-driven decision making reshape organizational politics. Drawing on the core insight that numbers lose their authority once mobilized in political contestation, one might expect that norms of data-driven decision making transform goal conflicts into battles over "what the data show." Yet data rarely settles such disputes. On the contrary, in the absence of hierarchical tie-breakers and amid competing methodological paradigms, debates often spiral into circular attempts to "debunk the debunker" (Kaplan, 2008, p. 741). In other words, arguments escalate beyond specific interpretations or data quality to challenge what counts as relevant evidence in the first place (Christin, 2020; Rerup & Zbaracki, 2021). Furthermore, even if subgroups were to agree on a shared standard of evidence (S. Feldman, 2004), the push for clarity might just as easily sharpen goal conflicts as resolve them. When interests diverge, ambiguity can be politically useful: it enables coalition building by allowing different groups to project their own priorities onto a shared initiative (Leifer, 1991; Reinecke & Donaghey, 2025). The clarity associated with numbers, by contrast, is more likely to expose and amplify underlying tensions.

This raises an important theoretical puzzle: coalition building requires either traditional authority mechanisms or authoritative evidence, yet regimes of data-driven decision making provide neither. They deemphasize formal hierarchy in the name of objective evidence but also subject data to political contestation, which erodes its authority. Existing theory cannot explain how coalitions form and persist under these conditions. This gap matters not only because data-driven decision making has become a widespread aspiration (Brynjolfsson et al., 2021), but also because failed coalition building is an independent source of the incrementalism that the literature on data-driven decision making has tried to explain with reference to myopia (Allen &

McDonald, 2025). This motivates our research questions: *"How do organizational actors build coalitions in data-driven organizations when arguments from data are inherently contestable and clarity is not always desirable? And under what conditions do these strategies fail?*

Methods

Research Setting

To develop a grounded answer to our research question, we analyze data from ethnographic fieldwork conducted by one of the authors while embedded in an organization operating one of the world's largest digital platforms for hiring expert contract labor, which we call "iLabor" (a pseudonym). Headquartered in the San Francisco Bay Area, iLabor[1] operates a digital platform that connects millions of freelancers and clients worldwide for remote, project-based work in areas such as software development, creative and design, and finance and accounting. Its 400+ full-time staff and thousands of external contractors were organized into core functional areas including marketing, sales, product, engineering, and operations. In addition to reporting to corporate executives within these domains, they also reported to initiative owners. Namely, the organization operated with a matrix structure, assigning staff from different functional areas to cross-functional strategic initiatives, such as Project MATCH.

During fieldwork, iLabor began overhauling its approach to matching freelancers[2] and clients. The original system gave freelancers and clients significant control over how they presented and searched for expertise. Search algorithms mostly relied on matching keywords to unstructured data from job posts and profile descriptions. As freelancer specialization deepened and clients'

[1] We selected iLabor based on a scoping study and expert consultations highlighting its significance. Reports of notable developments at iLabor made it a compelling case for deeper ethnographic investigation.
[2] We refer to the independent workers on the platform as "freelancers," reflecting the language used by our informants, though client companies often classify them legally as independent contractors.

demands grew more complex, this system created several challenges. First, the search algorithms were unable to match clients and contractors who used different terminology to describe the same expertise (e.g., "jurisprudence" versus "legal writing"). Second, freelancers who listed multiple skills or expertise areas on their profiles (e.g., mobile app development and graphic design) got unintentionally deprioritized in the search results for either skill. Third, the lack of structured data constrained iLabor's data science team's ability to understand the structure of supply and demand across market segments, limiting their capacity to "see the market," identify what was on offer, and pinpoint where shortages existed.

In spring 2017, Project MATCH was launched to address these matching challenges. Born from the 10% time of a handful of UX designers, product managers, and data scientists, it sought to replace unstructured, keyword-based search with a structured taxonomy that could classify freelancer profiles and client job posts more precisely. This entailed developing a comprehensive classification system, a taxonomy[3], for iLabor's marketplace. During 2017, this ambitious effort touched upon practically all parts of the platform's architecture and involved 80+ designers[4] (see Table 1). Together, they had to build the taxonomy, redesign practically all user interfaces for collecting and displaying information, and reconfigure governance systems to get freelancers and clients to use their new approach to data logging, algorithmic matching, and search.

== Insert Table 1 about here ==

The design work closely followed the logic of data-driven decision making. As is typical in larger cross-functional projects, the work was divided into six workstreams and coordinated by

[3] With transaction platforms such as Amazon, eBay, and Etsy, taxonomies are powerful digital infrastructures - arguably more powerful than reputations and rankings, which have received the lion share of attention - because they define the various market segments, i.e., who competes with whom and on what terms as well as who does not.
[4] We use the term 'designers' to encompass the various occupational groups involved in the design of iLabor's platform architecture, including UX researchers, designers, engineers, data scientists, product managers, and marketing managers.

product managers, with support from a technical program manager, and managers of functional units. Table 2 details the specific focus of each workstream and the occupational groups responsible for their execution. The workstreams produced different components of Project MATCH that would assemble into a minimum viable product (MVP), a working taxonomy system for one market category, which could then be tested in the market. Within workstreams, employees also ran experiments and collected data to inform decisions. In short, multiple development cycles were nested into each other, each iterating between development and testing.

== *Insert Table 2 about here* ==

iLabor's Project MATCH provided the ideal setting to investigate coalition building in data-driven organizations. First, iLabor's culture placed a strong emphasis on data-driven decision making, with experimentation deeply embedded in product development. This made it a fertile fieldsite for examining how data and experimentation were mobilized to justify decisions, forge coalitions, and shape the direction of design work. Second, Project MATCH represented a major structural change initiative, involving six interdependent workstreams and multiple occupational groups. With formal authority limited by iLabor's matrix-like structure, managers had to navigate goal conflicts through persuasion rather than exercising formal authority. Third, the research benefited from unusually deep access to the project as it unfolded. Close observation of collaboration and decision making in meetings, combined with in-depth interviews, allowed for detailed insight into the relationship between the way data was produced and analyzed, on the one hand, and mobilized for coalition building, on the other. That is, it allowed close insights about the link between the 'rhetoric and reality' (Zbaracki, 1998) of data-driven decision making.

Data Collection

We draw on data collected by one of the authors (henceforth field researcher) during ethnographic fieldwork conducted in 2017 and early 2018 (see Table 3). This method is well suited both for gathering data on work practices from the perspective of informants (Barley & Kunda, 2001) and for observing decision-making processes as they unfold in situ (Van Maanen, 2011). Access at iLabor was obtained following an academic referral, preliminary meetings, and the negotiation of a research agreement. The field researcher was introduced to employees as a researcher studying decision making in platform design and was provided with a company account granting access to internal systems relevant to the projects studied. Fieldwork was conducted over 15 months, including eight months of immersive onsite engagement. The study followed an open-ended and inductive approach, informed by a broad interest in how decision making unfolds in the design of platform technologies. While this initial focus was exploratory, the role of data in shaping decisions quickly surfaced as a central theme during fieldwork as informants repeatedly emphasized the importance of data and 'ground truth' within the organization.

During periods of onsite data collection, the field researcher spent roughly 60 hours per week at iLabor. Research techniques included onsite and remote observations, informal conversations, semistructured interviews, photography, as well as gathering work materials and online and archival documents. Reflecting the long hours made by Bay Area designers[5] and the distributed nature of the workforce, fieldwork days often began at 6 or 7 AM and ended in the early evening. Onsite fieldwork covered three phases (one month, three months, and four months), totaling approximately 1920 hours. Between these periods, the field researcher balanced data collection (e.g., remotely attending key meetings, following online exchanges, and conducting interviews)

[5] We use "designers" to refer broadly to all roles involved in new product design, including product managers, engineers, UX designers, and data scientists, who contributed to shaping platform features.

with theoretical iteration and data analysis. Observations and conversations were documented through audio recordings and fieldnote software, with evenings and weekends devoted to expanding notes, reviewing relevant work materials, and setting the focus of next day's fieldwork.

== Insert Table 3 about here ==

Observation. The field researcher spent eight months at iLabor's premises doing participant observation five days a week, although she was more of an observer than a participant due to the design work directly affecting the platform's core architecture. She had access to all areas of the organization, attended new employee introduction days, all staff gatherings, manager trainings, a women's summit, offsite events and strategic roadmapping sessions, product meetings and retrospectives, and a one-week annual engineering meetup. In the context of Project MATCH, the field researcher spent time with every designer across all workstreams. Due to the distributed nature of iLabor's workforce, work was largely organized and coordinated through meetings. As such, meetings became a central site of observation. The researcher shadowed project members by attending a wide range of meetings, including those with their managers, and which offered insight into how work was carried out and how design discussions unfolded in real time. Access to members' work calendars facilitated the selection of relevant meetings to observe.

In total, she observed 197 design-related meetings, including project, unit, and supervisory meetings, design sprints, internal demos, team debriefs, cross-functional meetings, roadmap reviews, as well as sessions with external consultants. During meetings, she took fieldnotes, took screenshots of work materials, and recorded 136 hours of audio (later transcribed), documenting how project members collaborated, made sense of evolving project demands, and navigated divergent interests across multiple workstreams. These observations further revealed the language,

data, and metrics groups used to construct shared understandings of their work. The cross-functional meetings proved particularly valuable for observing how decisions got made and project leads' efforts to resolve conflict by appealing to a shared ideology of experimentation.

Interviews. The field researcher conducted 107 semi-structured and more open-ended interviews, totaling 86 hours, with 54 project members from all key occupational groups, including product managers (26), UX designers (14), engineers (front- and back-end; 30), data scientists (6), data analysts (4), product marketing specialists (3), and customer researchers (10), as well as with iLabor's leadership team (14). Most people were interviewed multiple times during fieldwork. Interviews, typically lasting between 30 and 60 minutes, explored a range of themes including organizational roles and work practices, decision making and coordination, the use of data, metrics and technologies, alignment around product goals, and cross-team collaboration and tensions (see Appendix A). These interviews provided insight into how role-based values and cognitive frames shaped members' interpretations of the project. We used these interviews to uncover how different groups envisioned the future of the platform, reflected on the tradeoffs inherent to structural reforms like Project MATCH, and narrated both the authority and the limits of different sources of data. They helped us to understand both the broadly held beliefs about data and experimentation at iLabor and how members viewed their role in shaping decision making, resolving internal tensions, and aligning interests across roles and workstreams within the project. Informant names in the findings are aliases that preserve the cultural background of the original names.

Work materials, online, and archival data. We also gathered numerous documents throughout the fieldwork, including internal presentations, reports, product specifications, Jira tickets,

email threads, Slack messages, design mock-ups, work calendars, and engineering documentation. These artifacts offered fine-grained detail on how Project MATCH was scoped, debated, and implemented. In our analysis, we used a selection of 694 documents that were generated during project development to trace how principles of data-driven decision making were embedded in the development process, how assumptions about testability were formalized or challenged, and how communication infrastructures shaped the interpretation of evidence across workstreams. In addition, we reviewed 87 communications sent to iLabor users between 2017 and 2020, including onboarding messages, product announcements, and update notifications. They offered a lens into how internal stakeholders translated technical or experimental decisions into promises to users. Taken together, these multiple sources of data revealed how data-driven decision making at iLabor operated in practice, and how project leads resolved goal conflicts and other disagreements to forge a stable coalition for Project MATCH.

Data Analysis

To address our research question, we followed a grounded theory approach, iterating between puzzles emerging from the field and successive rounds of coding, while remaining in close dialogue with relevant literature (Glaser & Strauss, 1967; Strauss & Corbin, 2007). Our focus on how project leads invoked the ideology of experimentation to stabilize coalitions was not predefined but emerged gradually during data collection and analysis. This process broadly unfolded in four steps, detailed below.

Difficulties to build and maintain coalitions appeared early during fieldwork inside iLabor, where project leads described the difficulty of implementing consistent, structural changes to the platform architecture that would have a lasting market impact. This challenge was widely recognized across product and engineering teams and frequently lamented by informants, who voiced

frustration with the limits of piecemeal updates and a desire for more transformative change. As we gathered data on Project MATCH, we increasingly focused on the difficulties of building coalitions.

Establishing analytical frame. First, we went through interview and meeting transcripts and wrote descriptive memos to capture what was discussed in each setting. This helped establish a shared understanding of how Project MATCH unfolded over time. In parallel, we wrote analytical memos, examining how design work was organized in practice, focusing on the different roles and occupational groups involved. For each role (e.g., product, back-end engineering, UX design) we analyzed (1) guiding values/ideology, (2) ways to see the marketplace, (3) ways to intervene in the marketplace, (4) work practices, (5) decision making, (6) ways to advance one's agenda in the project, and (7) relations with other roles/groups. These memos formed the foundation of our initial coding scheme. Early codes for the UX designer role for instance included "respect users' emotions," "representing FL as best as possible in search," "market as visual object for users to interact with," and "iterative intervention informed by user research and digital trace data." Once we had established a shared analytical frame and initial coding structure, we imported the fieldwork data into NVivo, marking the transition to the next stage of analysis.

Coding. Second, we began systematically coding the corpus of fieldwork data, generating hundreds of first-order codes, which we then grouped into second-order themes derived in part from our analytical memos. Examples of additional themes included "how the taxonomy works," "technical requirements," "work constraints," "purpose and goals," "what to focus on (right now)," and "controversies and conflicts." Our initial coding concentrated on differences between roles, and we continued to write descriptive memos alongside this work. Their ongoing discussions were synthesized in four core analytical memos that examined (1) iLabor's organizational

commitment to what we referred to as "the engineering mindset," (2) how the organization of workstreams led to inconsistencies, (3) differences in how workstreams, thought about user autonomy, and (4) why problems remained unresolved and the desire for data-driven decision making. This revealed that alignment among actors was shaped less by formal roles or titles than by the parts of the platform they worked on (see Figure 1 and elaborated further in the Findings). These divisions shaped not only actors' goals but also how they "saw" the market, what they valued, what forms of data they trusted, and how they interpreted evidence. It also became clear that persistent conflicts of interests between groups lay at the heart of difficulties in moving the project forward. Accordingly, the next stage of analysis focused on cross-functional meetings where these differences became visible in specific goal conflicts, and we began tracing how project leads attempted to navigate and resolve them.

== Insert Figure 1 about here ==

Analyzing goal conflicts. Third, to deepen our analysis, we closely examined our fieldwork data to identify moments where goal conflicts became visibly apparent. We identified eleven recurring design issues where goals between different groups were clearly misaligned. For each, we analyzed how various groups positioned themselves and how data played a role in decision making. As it became evident that persistent goal conflicts were central to the project's stalled progress, we turned our attention to how project leads attempted to manage these tensions.

Coming from the literature on data-driven decision making, we noted that project leads were faced with a surprising problem: the need to 'let the data speak' often made resolution of goal conflicts more difficult, not less. This contradicted the assumption in prior research that objective evidence can provide a stable arbiter for decisions (e.g. Brynjolfsson et al., 2011). In our case, appeals to data created problems: it either antagonized subgroups or led to contestation.

The literature on quantification explained why: metrics are most authoritative when they coordinate organizational activities. When they are mobilized in political negotiations where distinct interests are at stake, they quickly become subject to contestation. Despite not offering clear intuitions about how this problem could be solved, the literature directed our attention to the question how the authority of numbers is constructed.

At this point, we realized that project leads knew that specific data rarely forged compromises and tried to sidestep the issue by appealing to the *ideology* of experimentation in more aspirational terms. In subsequent analysis, we therefore identified and coded three recurring strategies (see Table 4) that invoked this ideology ("assert experimental authority," "defer pending evidence," and "promise future resolution") and analyzed how they were used to settle specific goal conflicts.

== Insert Table 4 about here ==

Understanding epistemic fragility. Fourth, shifting to a more temporal analysis, we identified the MVP launch as a turning point in Project MATCH—one that undermined the belief in the ideology of experimentation. While strategies invoking the experimental ideal proved effective before the MVP's release (mid 2017), they rapidly lost traction afterward (late 2017-early 2018). In this later period, front-end designers gained influence by drawing on user feedback and qualitative methods. To better understand why the political appeals to the ideology no longer worked, we deepened the analysis of actual experimental practice.

We found that experiments at iLabor rarely adhered to scientific standards. Assumptions of isolation, control, and causal attribution were routinely violated due to user reactivity and entanglement with legacy infrastructure. In Project MATCH, these violations had been present all along, but they first became salient when multiple workstreams looked to MVP data to answer

central questions (centrality) and when the results themselves proved inconclusive (ambiguity). The increased scrutiny of experimental practices revealed that the MVP hardly counted as a legitimate experiment. This exposed a gap between the ideology of experimentation and its implementation in the context of iLabor. As a result, the symbolic power of experimentation weakened and with it the political strategies that built on it.

Findings

We present the findings in four sections. First, we introduce the political dynamics of Project MATCH and identify two core difficulties imposed by the need to resolve goal conflicts through persuasion with data: arguments from data were contested or yielded decisions that alienated parts of the coalition. Second, we introduce our theoretical model (summarized in Figure 2): project leads sustained Project MATCH not by resolving goal conflicts with data, but by appealing to a shared ideology of experimentation. We identify three strategies to resolve conflicts and keep the coalition together. Third and fourth, we trace how these political strategies relate to actual practices of experimentation. When experiments addressed central questions but produced ambiguous results, widespread scrutiny exposed methodological flaws and weakened the analogy between laboratory and corporation. This, in turn, undermined the faith in the ideology and the strategies that appealed to it. We call this dynamic the *epistemic fragility of experimentation.*

== Insert Figure 2 about here ==

The Challenges of Coalition Building in Data-Driven Organizations

Project leads at iLabor were political actors who had to build and sustain coalitions to advance their projects. Securing buy-in from management was not enough; project leads still had to persuade contributors to invest time and attention. With flat hierarchies and fluid roles, employees

had discretion over where to dedicate effort in a resource-constrained environment where staff split time across multiple projects. As David noted, "we traditionally bite off more than we can actually deliver on. . . so it's really hard for the company to rally around [any particular project]" (20170731-I-David). Team members could therefore quietly disengage from projects they did not like, stalling initiatives. David recalled that one project "had a beg, borrow, and steal kind of model to get any engineering support. And it slowed them down, it really limited what they're able to do" (20170728-I-David). Given these constraints, project leads had to continually ensure that the team was on board with the project.

Securing buy-in was not always trivial because employees working on different parts of the platform had distinct interests (see Figure 1). The platform architecture consisted of three principal components: the front-end, the back-end, and the governance system. The front-end included all user-facing elements, such as freelancer profiles, job posts, search pages, and the help center. Because front-end staff designed for end users, they prioritized user satisfaction and followed user-centered design principles such as "users should not have to wonder whether different words, situations, or actions mean the same thing."[6] Their performance metrics reinforced this orientation: they were evaluated on user satisfaction and sought to create an environment in which freelancers and clients could thrive.

The back-end comprised the systems that made the front-end work: hardware, databases, and algorithms that carried information between components. Back-end engineers' primary goal was system integrity over time. They viewed the platform as a complex, evolving architecture in which each design choice had long-term implications for the rest of the system. As one engineer

[6] 2017_11_29_PROD_categories_Search ATS Review "UX Quality Standards Checklist."

explained, the challenge was "to understand how the whole system will work to satisfy our existing architecture and how it should satisfy future usage" (20170703-I-Oleh).

The governance system included dashboards and administrative tools for managing the marketplace day to day. Market management staff utilized these tools to optimize for aggregate "ecosystem health," tracking key performance indicators such as fill rate, retention and growth, time-to-fill, and trust-and-safety outcomes. This group was most interested in the company business case: identify the optimal combination of matches that maximizes user growth and revenue.

In Project MATCH, members of these three groups worked in distinct workstreams (see Table 2) but interacted frequently in cross-functional meetings. Because their primary interests diverged (user experience for the front-end, system integrity for the back-end, and market balance for governance), design discussions often surfaced deep goal conflicts. Project Leads had to manage and resolve these goal conflicts to keep everyone invested in the project. This was the task of building and sustaining the coalition for Project MATCH.

Formal authority alone was rarely sufficient to ensure buy-in. Even in the more hierarchical engineering teams, employees were accustomed to resisting decisions that violated their core interests. For instance, early in Project MATCH, back-end engineers opposed product managers' efforts to limit the scope of work, insisting on building a full, scalable database system rather than an MVP. As one engineer explained, "business guys think 'you [engineering] guys start some work. And then, when we finalize our requirements, you just adjust the product.' But this is not working" (20170711-I-Danylo). Bringing them back on board required project leads to explain how exactly the MVP would scale later on, highlighting that progress depended less on formal authority than on acts of persuasion.

The need to persuade with data. Given that data lay at the heart of its business model, iLabor was deeply committed to data-driven decision making. The platform's central promise to freelancers and clients was the delivery of superior matches, an objective that most employees framed as a problem of measurement. Identifying the "perfect match" was widely understood to depend on collecting and analyzing high-quality data with the right instruments. A common refrain captured the positivist ideology at the heart of the organization's culture: "The more data we will have, the more we will be able to understand user behavior from both sides, from freelanc[ers] and from clients" (20171218-I-Theo). Similarly, UX designer Heath explained: "I think [it all] depends on the data you're getting. If clients gave you crystal clear data, and freelancers gave you crystal clear data, then you can improve the relevance [of matches perfectly]" (20171121-I-Heath).

Achieving such "crystal clear" data, they believed, depended on good design. Users submitted accurate and useful information only when the platform's infrastructure incentivized them to do so. In turn, whether a design decision was deemed effective often hinged on the quality of data it generated. Product manager Soren captured this circular logic: "I think if you [design interfaces] really well, then you've got really high-quality demand-supply reads that we can then use to identify where there's issues and opportunities in the marketplace" (20170706-I-Soren). Because data were viewed as iLabor's core competitive advantage, platform design work focused heavily on building infrastructures that could produce, capture, and analyze data about user behavior.

This emphasis reinforced the belief that design decisions should themselves be grounded in data. As data scientist Alice put it: "It's really a data-driven company. So, when something is wrong or broken, the analysts are using the data to [figure out what went wrong and] affect the business decisions that are being made" (20170629-I-Alice). This logic shaped the structure of platform design work. The workstreams of different projects followed principles of agile development that

emphasized speed and responsiveness to new data. "It's a part of our DNA," David emphasized, "it's being able to move quickly, but also change direction quickly" (20170628-I-David). Teams were expected to begin work before all parameters were fully known. "Something that we do [is to] fail fast," explained back-end engineer Homer. "This is a very important principle" (20170709-I-Homer). Failing fast meant running A/B tests, gathering user feedback, and conducting internal testing. Nested within the broader cycle of MVP development were smaller cycles of experimental component design. In short, data-driven decision making was central not only to product outcomes, but to how the organization imagined, coordinated, and evaluated work. When disagreements emerged, employees expected a principled argument from data to resolve the issue.

Political Problems of Data-Driven Decision Making. This emphasis on data as the arbiter of decisions created distinct problems for project leads who tried to maintain the broad coalition required for Project MATCH. While all groups liked the idea of a smooth and integrated system that would produce better matches, the project required substantial changes to practically all parts of the platform and threatened to limit users' autonomy in decisive ways. Front-end roles therefore clashed with market management and back-end when specific design decisions threatened freelancers' or clients' ability to act as they had in the past. Conversely, back-end engineers were wary of fundamental changes that were not planned well. In particular, they worried that MVP-stage stopgaps would not be resolved later on, "we've had a lot of technical debt built up [over time]" (20170719-I-Rahul). Furthermore, there was a fundamental conflict over whether to prioritize good matches for clients or opportunities for freelancers: "we're very client focused . . . but we won't be able to attract those clients if we don't have good freelancers" (20171121-I-Celine). As a senior product manager framed it, the core divide was whether freelancer talent was a commodity or an asset to acquire and retain (20171120-I-Harper). Even seemingly technical choices frequently

touched on these issues and could quickly become highly contentious, threatening to derail the coalition.

These kinds of goal conflicts proved difficult to resolve with reference to data. Project leads faced two problems in particular. First, data were not a neutral resource that could simply be mobilized to resolve disagreements. Groups drew on different methods shaped by their responsibilities: front-end teams relied on interviews, stylized personas, and user studies; back-end engineers preferred technical metrics and performance testing; market managers used statistical models and panel data to track aggregate trends. As product manager Zane put it:

> Although we pretend to be a data-driven organization, we're not really, because we're not all looking at the same data. And we're not all judging it the same way. And we don't even have all the same objectives […] It's just that you armor yourself with a bit of arithmetic, in order to make a case for something that you want to do anyway (20171130-I-Zane).

An early disagreement between product managers and back-end engineers illustrates the point. The issue was when to "deprecate" a specialization, i.e., retire it from the taxonomy once it was no longer in use. Product managers argued from "use cases," pointing to terms that had fallen out of circulation. Back-end engineers, by contrast, worried about the structural impact on the database. As their team lead, Rahul, pointed out, to deprecate a specialization could orphan existing data entries. He proposed graph-theoretical measures to diagnose "entropy" and "independent edges" in the database (2017-06-29-O-Back-End Engineering-Taxonomy). Each group thus privileged different forms of evidence because they conceptualized the problem differently. What counted as "decisive" data for one audience carried little authority for another, making it difficult to resolve a conflict with data.

Second, even decisive data could destabilize rather than sustain the coalition. When data pointed clearly in one direction, they forced project leads to side with one party to a conflict and thereby frustrated the other, prompting disengagement. Early in Project MATCH, back-end and

front-end groups clashed over how tightly the new freelancer profiles should be integrated with the rest of the system. Front-end resisted forcing freelancers to translate their job histories into the new taxonomy, because they feared users might find the work onerous and constraining. Back-end engineers objected that leniency would create inconsistencies. Substantial evidence that users were unwilling to do the work initially seemed to decide the issue. But engineers then feared the risk of the resulting inconsistencies and quietly withdrew support. Rather than develop the requested structures, they kept the taxonomy in an Excel sheet outside the system. This work-around protected iLabor's database structure but made scaling and integration with other components of the project practically impossible. Even when it was clear what the data showed, this clarity often did not help to sustain the coalition between the three groups.

In short, the problem was not just that data were ambiguous or that evidence was contested. It was that any data-based resolution would reveal and amplify goal conflicts that expressed incompatible interests. Project leads needed a way to appeal to data without inviting contestation and to obscure rather than surface goal conflicts, a recourse to authority and strategic ambiguity.

Invoking the Ideology of Experimentation

During fieldwork, we observed that project leads rarely drew on specific analysis or data to resolve disagreements and keep the coalition together. Instead, they appealed to the shared belief in the scientific method. Specifically, they relied on three strategies: asserting experimental authority, deferring pending evidence, and promising future resolution, all of which invoked the ideology of experimentation as the gold standard for data-driven decision making. While asserting experimental authority aimed to end debates over the best course of action, deferring pending evidence and promising future resolution were used to suspend irreconcilable goal conflicts in order to maintain coalitional buy-in and project momentum (see Figure 2). Table 5 summarizes

the 11 key design conflicts, how they mapped onto the interests of the three groups, and how the three strategies were used over time, with varying degrees of success.

== *Insert Table 5 about here* ==

Asserting experimental authority. Project leads' first strategy to resolve conflicts was to force decisions with an appeal to authority. They did not exercise the authority of their role in the organization or their occupational expertise. Instead, they discounted unfavored positions as not backed by experimental data, or they declared a preferred position to be the result of an experiment, even when it was not. In each case, they appealed to the authority of the experimental method, rather than the substance of an argument or finding.

Although iLabor lacked a unified methodology for evaluating evidence and different groups preferred different sources of data, a clear status hierarchy existed among ways of knowing. Experimentation or formalized "testing" was the de facto gold standard, not only because it was foundational to the development cycle at iLabor, but because it was seen as the most scientific way to "know" the market. Those who worked with qualitative data often expressed resigned frustration at the blind belief in "scientific" methods. As designer Audrey observed:

> "I don't think that the conclusions drawn from [experimental] data are that deeply questioned. You are questioned if you don't have those kinds of numbers to back up your statement. But there's not a lot of questioning of the logic of how those numbers were arrived at or interpreting the data." (20171122-I-Audrey)

The power of this hierarchy also became visible when it was violated. In one interview, data analyst Scott described his frustration when a statistically significant finding from a quasi-experimental study was ignored: "If we have more experienced freelancers active we fill more posts. . . Even if you control for the number of posts, still, adding additional applicants means additional jobs get filled." Despite this, "the highest-level people are pushing back against that." He added, "the data was intuitive" and that this "is literally an economics 101 issue, because that's how you

understand supply and demand" (20171129-I-Scott). This commonly accepted status hierarchy thus enabled a strategy to bypass disagreements: rather than discuss the merits of any particular set of data, project leads could discredit positions not based on "truly" scientific data or boost their own by claiming that they were.

One use of this strategy took the shape of academic critique. For instance, front-end engineer Theo drew on a standard point from experimental psychology when he challenged the validity of qualitative interviews: "Nobody can guarantee that [a freelancer] is able to accurately describe his skills, his profile. . .This is even worse for the client, to be honest. [Nobody is able to guarantee] that he's able to describe the job that he needs done, what he is looking for" (20171218-I-Theo). A similar invocation of positivism was directed at design with the aid of "personas." During a lunch meeting of the project's core team, Zane offered a textbook objection:

> "We create idealized perspectives on customers. And then I think, especially if you misuse a persona, which is the easiest thing to do, you just say, like Bob, the small business owner, would never do this. And because now you've got this mental model of your persona, you're not actually testing your design against a real use case." (20170616-O-PRODCORE)

Here, the persona is described as epistemically weak in comparison to experimental forms of knowing: it is resistant to falsification and detached from empirical testing and should therefore be discounted. These kinds of dismissals were effective because the deep commitment to data-driven decision making rested on a valorization of "objective knowledge" from scientific experimentation: the broadly shared ideology at iLabor.

This strategy was also used in its affirmative form, effectively claiming that a position had experimental backing even though the actual evidence was informal or correlational. In September 2017, as part of an annual week-long offsite, the company convened all product managers, data scientists, and engineers working on Project MATCH to strengthen collaboration and build momentum amid areas of non-alignment and slower-than-expected progress. During one meeting,

an engineer objected to a proposed database structure for the taxonomy, calling it rigid and hard to manage. The founding team responded by invoking a feature called Talent Spec. As Zane recalled in an orientation meeting: "I've spent five years trying to fix the search system. And one of the ways we found that really works is 'Talent Spec,'" a tool that allowed clients to specify in standardized language what kind of talent they were looking for (20170914-O-ENG_OFFSITE_1). The sense that "it really works" did not come from a formal experiment; it was based on post-launch analysis and qualitative exit interviews. Yet in the meeting, product manager Tuan presented it as the outcome of a structured test:

> "What gives us confidence in this thinking is that in June . . . we rolled out a test on the Freelancer search side, called 'Talent Spec,' . . . And we saw the numbers go up, all the numbers go up in the [search] funnel." (20170914-O-ENG_OFFSITE_1)

The claim that an experiment had revealed a meaningful impact was enough to silence further objection. The manager never discussed any details, and no one in the room revisited the actual analysis. Instead, Tuan simply stipulated that the position could not be doubted. The authority of experimentation, invoked in form if not in substance, ended the discussion. We observed similar dynamics in conflicts 6 and 9 (see Table 5), where the strategy of asserting experimental authority was used to undermine dissent or end debate of a contested decision.

Deferring pending evidence. A second strategy to stabilize the coalition was designed to suspend conflicts over irreconcilable differences between groups. Rather than try to decide a contentious issue, project leads often argued there was not enough good data to determine which side was right. Only a rigorous test could produce this data, but this first required a working prototype. Accordingly, everyone had to put their differences aside and pull together to build it. Unlike direct invocation of experimental authority, which shut down disagreement, the deferral strategy used ambiguity to put irreconcilable goal conflicts on hold, preserving the coalition by deferring resolution to a future moment of empirical clarity.

During an early product review of the mobile flow for creating specialized freelancer profiles, a disagreement surfaced between the leads of Project MATCH, who wanted to encourage users to complete the profile as thoroughly as possible, and UX designers, who were concerned that users might feel trapped in the flow of interfaces for profile creation. At that point in development, the interface did not allow users to exit to another part of the website immediately. A product manager defended this design choice: "You're really immersed in this experience, and you can't sort of wander off to your messages or look at your existing profile or anything else. I think that's a very intentional choice here." (20170705-O-DESIGN_mobile review) Pushing back, a UX designer proposed adding buttons that would allow users to preview their final profile and exit the flow more easily: "I think by including that preview, doorway, into this model, it then becomes a natural transition into the rest of the product . . . And maybe from that point, you could sort of exit the mobile experience and do something else." Rather than engage directly with the underlying question of whether the design should prioritize user autonomy or data logging, head of UX design David deflected. He proposed deferring the decision by testing both versions in the future:

> This is something we can easily test. You talked about a preview, so you sort of see how [your entries] fit into your service profile as you go along. And another step, which could precede each one of these, is to hit home the messaging around why you're doing this, set up that value proposition before they go on to the next step. This would add additional steps to the overall flow. But I think if the goal really is around quality, then I think that's something you'd want to test. (20170705-O-DESIGN_mobile review)

By allowing designers to incorporate the flexibility they wanted while postponing a final decision, David resolved the conflict without requiring consensus on the underlying issue. The meeting proceeded without further disagreement. Over time, deferral often led to design bloat: conflicting visions were encoded into the MVP with the promise that they would be tested later - but rarely were. We observed similar dynamics in conflicts 1, 2, 3, 4, and 5 (see Table 5), where the strategy was used to suspend irreconcilable goal conflicts and maintain buy-in.

Promising future resolution. Project leads used a related but distinct strategy when the MVP included features that clearly violated the goals of one group. With no active disagreement that could simply be deferred pending more data, they faced the challenge of reconciling groups to a decision that directly contradicted their interests. To do so, they declared the current design a provisional solution and promised that the experimental development cycles would ultimately deliver an ideal version that would render today's conflicts irrelevant.

When the core team pitched Project MATCH to the broader organization in May, they received enthusiastic support. Many were frustrated with fast, incremental work and eager to contribute to something more ambitious. As back-end engineer Yaronslav put it: "the dream of every engineer is to have an ideal situation when you're building a . . . project from scratch, [where] you do not need to support any legacy system or integrate with some other old parts of the system [and where] you can choose technologies by your own" (20170622-I-Yaronslav). Those involved in Project MATCH liked to imagine the future taxonomy system as a perfectly ordered mall, one in which freelancers and clients could effortlessly find one another because the architecture itself organized the space. As Tuan described it:

> "It's very similar to going to the mall. There is a place that has a wide array of stores and things to purchase. And I can go there to find what I'm looking for easily. Similarly, on the seller side, here's a place where I know I can be found, like my store is fixed. People can find me easily and they can understand what it is I have inside my store." (20171205-I-Tuan)

In this future-oriented vision, the different interests of front-end, back-end, and market-governance roles would all be resolved: the system would be more restrictive, yet users would express themselves more easily than before; data would be saved and retrieved through a single, unified, and internally consistent structure; and matching rates would improve.

Promising future resolution appealed to this idealized future. Because experimentation would yield iterative perfection, current design flaws appeared temporary and unimportant. Concerns

were framed as "in-between state losses" (20171205_O_E2E review FL search), described as short-term setbacks that would resolve as the system matured.

A telling example of this strategy arose in discussions with engineers responsible for the database requirements. They had to develop new data structures for the taxonomy system and determine how to relate the new categories to existing freelancer profiles and job posts. Product managers suggested the creation of a "general profile" that preserved legacy profile data. Product manager Celine explained: "this is basically where the Freelancer's existing profile is going to live. And that's really so that the Freelancer isn't losing their old profile. It's still there if they want to view it or the client wants to view it" (20170615-O-ENG-CFD). This proposal immediately raised red flags for engineers, who believed that this would require them to preserve two inconsistent data architectures. As engineering lead Denys put it: "That sounds like we need to support the old back-end and the new back-end by hand. It might be very problematic." Heath tried to reassure him: "To be clear, we should migrate all of the old profile data into the new profile data. Although displaying the old profile data, it will be on the new service [profile]." But Denys remained unconvinced. From his perspective, the old profile data was incompatible with the new system. He tried again, suggesting that the old profile data "does not exist as general profile data in the new world. So, for the new profile, you'd have to support new services but also old [if you wanted to keep this old data]." His concern was that keeping the general profile active would create inconsistencies: two parallel tables with incompatible data structures.

At this point, Heath stepped in to reframe the issue: "Keep in mind that [the general profile] won't always exist. Once all the work history has been associated and we decide to cut off the old profile, that type of general profile will disappear." (20170615-O-ENG-CFD) Because Denys was already reasoning from the logic of full implementation, which required a single database for the

whole system, it became easy to present his concerns as real but temporary. Once users had updated their profiles and the new system had rolled out completely, the inconsistency would vanish. Maintaining both systems was thus framed as an unfortunate but short-lived inconvenience. Other engineers quickly aligned with this vision, proposing solutions to help nudge users away from their legacy profiles. One suggested: "when users add some services, like more than two, we can add a notification on basic profile, saying: 'looks like you already filled your services total information there. So, no need to keep this basic profile anymore, you can just close it.'" Celine confirmed: "ah, yeah, that's the goal for everybody eventually." (20170615-O-ENG-CFD)

Thus, promising future resolution leveraged the broad appeal of full implementation and belief in a data-driven process to achieve it. It allowed project leads to address concerns about user autonomy by invoking a future taxonomy that would perfectly reflect users' goals and preferences, while also soothing back-end concerns with the promise of a unified market infrastructure. Similar dynamics appeared in conflicts 6, 7, 8, 10, and 11 (see Table 5), where this strategy afforded ambiguity to mask unresolved tensions behind a shared vision of scientific progress.

However, these three strategies did not work consistently. As we show next, their impact shifted after the MVP was launched. By examining post-MVP dynamics and comparing them to those in another project, we show that the power of the ideology of experimentation hinges on the degree to which experimental practices remain isolated from scrutiny.

Epistemic Fragility: When the Ideology of Experimentation Unravels

While project leads initially succeeded in keeping the coalition together, their strategies began to falter after the MVP launch, and it became increasingly difficult to resolve goal conflicts between different subgroups. Eventually, front-end personnel withdrew support from project leads' vision and resisted any development that lowered users' autonomy, which undermined the core

premise of Project MATCH: that users would have to gradually express themselves in the terms of the taxonomy. To understand why appeals to the ideology of experimentation no longer worked, we now turn our attention to the reality of experimentation at iLabor.

We found no sudden decline in the quality of experimentation at iLabor; basic standards were frequently violated throughout the process.[7] What changed was not the methodological rigor of experiments, but how they were used and what kinds of answers they produced. As Figure 2 outlines, when experiments addressed (1) decisions of central importance to multiple stakeholders, and (2) produced ambiguous results, they triggered widespread scrutiny. This first revealed not only the methodological flaws of experimentation, but a deeper incompatibility between the ideal vision of pure science and the messy realities of product development. This, in turn, undermined the ideology of experimentation and the strategies that built on it.

Centrality. Before the MVP, experimentation mostly took the form of A/B testing elements of the new product design within individual workstreams. Designers used these tests to answer simple design questions (e.g., the color of a button). But following the MVP release, there was a general expectation that the launch would serve as an experiment to validate the central premise of the initiative: to increase search relevance by 10%.[8] Mid-November, iLabor first launched the functionality to create specialized profiles for several hundreds of freelancers working in mobile development. A corresponding "responsive job post" feature was also released to a randomized subset of clients likely to hire in this category.[9] Two weeks later, the user base was expanded to include over 10,000 freelancers and 50% of clients seeking to hire in the mobile development

[7] 20170912-O-OFFSITE; 20170710-MATCH-PROD-DA-doc Experimentation at iLabor Presentation.
[8] 20171121_MATCH_PROD_doc_Categories Success Metrics
[9]20171204_MATCH_PROD_doc_Specialized Profiles Launch Update.pdf; 20171205_MATCH_PROD_doc_Initial Launch Findings_ Job Post - Google Groups

category.[10] Mid-December, the new freelancer search and job search were launched to 10% of users, and this was increased to 40% after the Christmas break.[11]

Almost immediately, data began to stream in. “We’re watching a few of them in real time,” Zane shared with Harper during a key project leads meeting as they were observing life processes of profile creation in the event logs. “We're getting exactly what we're asking them to do. So, I think it's positive.” “Nice. It's exciting. It's out there!” Harper said enthusiastically (20171120-O-Category Leads). Once the new search functionality went live, data analysts received aggregate metrics about changes in match quality, the core outcome on which the entire project hinged. Internal announcements celebrated the rollout using the language of experimentation, stating, for example, that clients and freelancers had been “allocated to treatment or control groups” for the new features.[12] In short, the MVP launch was framed as an experiment that would soon answer the initiative’s core question. Yet, that initial optimism quickly gave way to confusion.

Ambiguity. The quality of search results did not depend solely on the taxonomy; it depended on how freelancers and clients *used* the new features. If they did not add relevant information to the specialized profiles, matches would remain poor. As Zane put it: “Match quality is conditioned on the quality of the profiles that are being produced by the system” (20171205-O-Initial Launch Findings). Unfortunately, only a small group of freelancers took advantage of the new possibilities immediately. By December 14th, only 1,054 of the more than 12,000 invited freelancers had published a specialized profile. Many had failed to select meaningful specializations or attributes that would differentiate their profiles. As Zane lamented: “Either people have only selected the

[10] 20171220_MATCH_doc_CustomerCommunications Freelancer Search with suggestions launching today to 10 of users
[11] 20180118_MATCH_O_Categories 2018 kickoff
[12] 20171205_categories_Initial Launch Findings_ Job Post - Google Groups

first couple of attributes that are mentioned . . . or when they click through, they just click everything. Check, check, check. Then they're not telling us anything!" Cleo added, "There are also a lot of people in the first batch who just copied their general profiles [into their specialized profile]. Sometimes they just didn't do anything at all!" (20171218-O-Working Group) As a result, "The overall focus score was 0.18, meaning that a little less than one out of five terms [on freelancers' profiles] were unrelated [to their specialization]." Though the focus scores improved by January, inconsistent use of the various MVP components still made the data difficult to interpret, even for the project's designers. It was unclear whether the observed 2.5% improvement in matching quality came from the new functionality, user behavior, or some combination of the two.

Scrutiny. While ambiguous findings often led to reevaluations and new experiments within workstreams, the MVP launch results triggered widespread scrutiny. Several meetings were held to discuss the results, and as the central question remained unresolved, people across workstreams began taking a closer look at the launch. Project members soon realized that calling the MVP an 'experiment' was misleading, given how it had been implemented. First, there was no valid sampling strategy. The initial treatment group had been randomly selected from freelancers with any earnings in the mobile development category. However, as Celine pointed out, "clients don't always put their jobs in the right category," meaning many freelancers in the first batch were misclassified and thus were never appropriate test subjects. "A lot of these people actually ended up being graphic designers, UX designers, and stuff," she explained. (20180118-O-Categories 2018 Kickoff). To compensate and expand the sample size, product managers had then manually added freelancers to the second batch, further reducing the consistency and integrity of the selection process.[13] Worse still, they had never defined an active control group for comparisons.

[13] 20171204_categories_Specialized Profiles Launch Update

The ideal unravels. Beyond these technical issues, project members also began to realize that the MVP violated core principles of experimental logic, calling into question the previously unexamined analogy between the ideal of 'hard science' and the messier realities of product development. Rather than serving as a discrete intervention in a stable environment, MATCH's MVP comprised multiple components that were deeply intertwined with iLabor's existing matching system. To ensure that freelancers in the treatment group remained competitive in iLabor's broader market, the MVP allowed them to maintain both a general and a specialized profile. An unintended consequence of this decision was that the new freelancer search algorithm had to integrate legacy keyword-based methods with the new taxonomy-based features.[14]

This had three important consequences: First, because clients could hire from either pool, freelancers in the treatment group were competing directly with those in the control group. Second, there was no clear baseline for evaluating the treatment. The system no longer maintained a clean separation between treated and untreated experiences. As a result, the mobile development category no longer offered a version that used only general profiles, making any comparison effectively contaminated. Third, freelancers in the treatment group had access to both general and specialized profiles, and used both. Accordingly, the experiment did not test the value of specialized profiles in isolation, but instead tested the combined use of both profile types within a system where that option was not available to everyone.

Together, these flaws made it impossible to understand and evaluate the MVP as an experiment. As data analyst Scott observed with some exasperation: "There are some things that you can't do. So if you can't measure what's happening because you can't [separate a control group],

[14] 20171220_MATCH_doc_CustomerCommunications Freelancer Search with suggestions launching today to 10 of users

there's no counterfactual" (20171129-I-Scott 1). The disillusionment also came out in customer researcher Melissa's statement that the different tests for Project MATCH "are not pure experiments, because people always want to kind of buff and polish around it. There's nothing to be done about that. It's just the way it is. It's software development and web development" (20171204-I-Melissa). These quotes reflect a growing realization that the analogy between the scientific laboratory and the corporate context of product development was flawed.

The strategies fail. As confidence waned in experimentation's ability to deliver clear answers, or in a full implementation that could resolve goal conflicts, strategies that relied on the authority of experimentation began to lose influence. This shift became particularly visible in a final discussion over launching the MVP's final part: the new freelancer search experience. Head of product marketing April raised a concern: freelancers' profiles would appear differently in search tiles than they had been led to expect. While freelancers had been told that their tiles would show their full range of skills, the new system displayed only the specialized profile most relevant to the client's query. This design reflected the intended logic of the taxonomy system: promoting specialization to improve relevance scores. But April, acting as an advocate for freelancer autonomy, cited customer research suggesting that users would be upset if their general profile became less visible, especially after product marketing had promised that the new system would not compromise their ability to showcase the breadth of their offerings. She explained:

> Let's say I have a general profile. It's like, I'm a mobile dev guru. And then I also have an iOS dev profile . . . I may say: Well, I specialize in mobile dev [not just iOS dev. Now] you're showing my iOS card, but the point was to let clients know that I also do general mobile dev work. (20171205-O-E2E Review of FL Search

April thought that all profiles communicate some specialization and that freelancers should not be restricted in how they advertise their skills to clients. Celine responded: "Yeah, we didn't consider the fact that they might want to highlight their general profile [in search]." At this point,

Heath joined in and warned that April's point contradicted the fundamental logic of Project MATCH, which was to specialize. But still, April continued to advocate for the user:

> Heath: The tricky thing is it's hard to tell, [or] highlight [on freelancers' search tile], what they do in their general profile *because it's everything*.
> April: I know, I know . . . But I think freelancers will call us on that, because we're not indicating in any way that they do other stuff.
> Heath: I think it's okay. We're just gonna have to figure out if that's a problem or not. The whole point of this system is [that] it's better to say that they specialize in iOS development than say, like 'hey, I specialize in all this other stuff, if the client is looking for specifically iOS development. (20171205-O-E2E Review of FL Search)

At this point, one might have expected this strategy of *promising future resolution*, to overcome resistance or at least to buy enough time to maintain alignment, to work. But pointing to the benefits that would later be derived from specialization was not enough. April held her ground and cited Celine's own prior communications with freelancers: "Ehm so Celine, I looked at the comms. You did say . . . that 'when you have multiple profiles, there'll be an indicator that lets clients know you have other profiles that can be viewed.'" Heath then became more explicit and tried to *defer* the concern by calling for more data: "It's kind of tricky until we see how clients interact with it. Do they pay attention to it? All we know now -" But April cut him off:

> "But I just want this to be something successful for freelancers as well. And it is something that when we originally scoped it out, we said we would have. . . I think that [this issue] has the potential to come up: hey, now you're showing search results, which is great, but the client doesn't have a full picture of my work." (20171205-O-E2E Review of FL Search)

Her strong reaction "I want this to be something successful for freelancers as well" reflects a turning point in her perspective, away from the belief that experimentation could surface and resolve critical issues. Instead, she urged the team to address this anticipated shortcoming now and address the primary interest of front-end staff: customer satisfaction.

A similar dynamic unfolded in discussions about whether teams should adjust the MVP to satisfy customers or develop a roadmap for fuller implementation. While the former threatened to integrate the MVP with the rest of the system, the latter would push ahead to realize the original

vision. In a working group meeting, Zane and other project leads discussed the best way forward. "If you've already launched," Zane stressed, "the priority is finding the market fit," i.e., making sure that the MVP worked for customers. Worried that a fuller implementation would be sacrificed to the concerns of front-end personnel, data scientist Thien jumped into the conversation and tried to assert experimental authority:

> Can I interrupt you guys? . . . So, so far, I think we've been thinking about the MVP as a test. And [given that] it is a study like this, there should be a session where we look at the results, [where] we want to assess the quality of the profile score. Maybe we want to look at the initial search results and see how this MVP helps us to get a better understanding of the freelancers and jobs, and also how to improve search, right? So I think these should be the next steps, like how we spend our time in the next two weeks, right? And hopefully, based on our findings, you can drive the roadmap, see what's actually important [for full implementation]."

Yet, his appeal fell flat. Noting the immense difficulties of treating the MVP as an experiment, program manager Chiyo replied with evident frustration: "So, Thien, it's going to take some time for us to get to that learning stage where we have enough data to figure it out and extract some learnings out of it. But in the meantime, we have to keep the flow continuing, right? We have to work on other stuff, we just cannot stop here." (20171204-O-Categories Working Group) In an earlier interview, UX researcher Cleo echoed this shift: "We don't really see [the effects of Project MATCH] until all the pieces come together . . . But in the meantime, we still want to have a good idea on what and how well we're doing on these intermediate things" (20171130-I-Cleo).

Hence, as confidence in the ideology of experimentation waned, the authority of its methods gave way to more pragmatic forms of evidence. With very small numbers of active users, insights came from qualitative interviews, session logs, and individual journey reviews, methods that normally carried less authority in the organization. Asserting experimental authority was no longer enough to win an argument, nor was reliance on less "scientific" data necessarily viewed as a weakness. Because project leads were no longer able to pull front-end designers on board with their strategies of deferring pending evidence and promise future resolution, the coalition

fractured, allowing front-end designers to push for an independent agenda that focused on short-term user satisfaction.

Counterfactual Analysis

When we analyzed the weakening of Project MATCH's coalition, we considered an alternative explanation: perhaps the MVP simply marked the point at which groups expected the promises of project leads to be realized. The coalition would then have collapsed because the experiment failed, not because faith in experimental ideology waned. To evaluate this possibility and stress test the two conditions of epistemic fragility (centrality and ambiguity), we conducted a counterfactual analysis with Project LOCAL, another initiative launched in 2017. Though more lightly covered in our data, the projects shared core features: cross-functional collaboration involving the same occupational groups, severe goal conflicts between front-end roles (who resisted restrictions on user autonomy) and market governance (who prioritized aggregate outcomes), MVP-based development, and significant implementation problems. Yet Project LOCAL's coalition remained stable throughout. Appendix B provides a detailed analysis showing that the MVP release itself did not undermine strategies of deferral and promise. Rather, the project kept experimental practices isolated from scrutiny by either keeping them peripheral or producing unambiguous answers to central questions, supporting our theory of epistemic fragility.

Discussion

Data-driven decision making is often understood as a set of tools and processes that reduce environmental uncertainty (Luca & Bazerman, 2021). But because epistemic technologies change how decisions are made, they also reshape organizational politics. This paper examined how project leads navigate a core dilemma: the need to justify decisions with appeals to 'what

the data show' collides with the need to build coalitions across divergent interests. Arguments from data not only invite contestation when they are meant to resolve goal conflicts (Espeland & Stevens, 1998; Levina & Orlikowski, 2009) but also constrain the strategic ambiguity often essential in early coalition building (Eisenberg, 1984; Reinecke & Donaghey, 2025).

We show that project leads sustained support not by resolving conflicts with specific data, but by appealing to the *ideology* of experimentation, the belief that structured testing yields objective insight. This ideology enabled three strategies: asserting experimental authority, deferring conflicts pending evidence, and promising future resolution. However, these strategies proved effective only while experimental practices remained peripheral and unexamined. We conceptualize this as *epistemic fragility*: the conditional authority of experimental ideology that collapses under the very scrutiny it purports to invite. When experiments addressed central questions but produced ambiguous results, scrutiny exposed methodological flaws and the gap between laboratory ideals and organizational realities. This undermined both the ideology and the political strategies that depended on it. Below, we discuss how these findings contribute to three distinct literatures.

Contributions to Understanding the Symbolic Power of Numbers

The paper opens new avenues for research on quantification by suggesting a distinction between the authority of specific metrics and the ideology of science that undergirds these metrics. Research on quantification shows that specific metrics gain authority through intensive construction and consecration: elaborate production, justification, and embedding in organizational routines (Espeland & Stevens, 1998; Mazmanian & Beckman, 2018; Stice-Lusvardi et al., 2024). Authority stems from attention: the more a metric's production and meaning are elaborated as core to the organizational process, the more powerful it becomes. We find that experimental ideology operates through the opposite process: it maintains authority precisely by keeping actual

practices peripheral and unexamined. While specific metrics require attention to gain power, experimental ideology requires inattention.

This inverse relationship rests on a distinctive feature of scientific ideology: its core value is contestation. Scientists earn credibility by subjecting claims to rigorous scrutiny and empirical falsification. Yet when this scrutiny is applied to organizational experiments, it quickly reveals that practices rarely conform to the ideal. While even laboratory science can deviate from positivist ideals (Pinch, 1993), this is especially true in corporations where quick decisions override methodological rigor (Luca & Edmondson, 2024; Marres & Stark, 2020). The more ardently scrutiny is applied, the more it reveals organizational experimentation to be flawed and inconsistent. Unlike other ideologies such as nationalism or American exceptionalism that are protected from scrutiny via tautology, appeals to scientific ideology can thus cannibalize their own authority. Once scrutiny turns inward, the ideal is quickly unmasked as normative fiction.

The strategies of deferral and promise function as mechanisms that shield ideology from internal contradiction. By projecting resolution into the future, they shift attention from current experimental practices toward future possibilities. They create epistemic distance, a gap between ideal and instantiation (Guala, 2005), that preserves authority by preventing examination. Epistemic distance differs from the "black boxing" identified in technological systems (Latour, 1987). Black boxes are stable inscriptions that resist interrogation by concealing complexity (Anthony, 2021). Epistemic distance is an active accomplishment maintained through temporal deferral. When deferral fails, the distance collapses, and ideology becomes vulnerable to scrutiny.

Contributions to Coalition Politics in Organizations

Our findings reveal a previously unrecognized mechanism for coalition building where traditional authority is weak. Prior research focuses on hierarchical settings where leaders control present resources, budget allocations, structural rearrangements, or hierarchical fiat to secure buy-in through bargaining or integrative structures (Cyert & March, 1963; DiBenigno, 2018; Levinthal & Pham, 2024; March & Simon, 1958; Mithani & O'Brien, 2021). This literature offers limited guidance when such authority is weak, a condition increasingly common in data-driven organizations that deemphasize hierarchy in favor of evidence-based decision making.

We find that project leads built coalitions by, on the one hand, appealing to the status hierarchy of methods, where positions could be elevated as truer than others if they appeared to be backed by experimental studies. On the other, project leads managed to transform the future into a resource. Strategies of deferral and promise worked by transforming present conflicts into future empirical questions, creating provisional zones of indifference (Barnard, 1938) where participants accepted decisions not because they agreed, but because they believed future data would vindicate their positions. This temporal authority differs fundamentally from traditional forms. Managers exercise hierarchical power by allocating resources in the present; professionals build expert authority through demonstrations of competence in 'scut work' (Huising, 2015). Temporal authority operates through deferred validation: members accepted deferral and promise because they believed experimentation would eventually reveal ground truth. This temporal dimension proved particularly robust within iterative development cycles. Because experimentation involved nested iterations, project leads could always point to the next iteration for clarity. The future perpetually receded, making temporal authority renewable rather than consumable.

Our findings also extend research on strategic ambiguity in coalition building (Eisenberg, 1984; Leifer, 1991; Reinecke & Donaghey, 2025). Prior work shows how vague language enables coalitions by allowing different groups to project preferred interpretations onto shared goals. We show that leaders achieve ambiguity through vague processes. By appealing to the authority of future experiments, leaders enable all parties to project preferred outcomes into an undetermined future. This process-based ambiguity is powerful because it promises eventual clarity, whereas linguistic ambiguity may seem evasive. Experimental ideology offers legitimated deferral: we are not avoiding conflict but waiting for rigorous testing to resolve it properly.

Contributions to Understanding Experimentation and Incrementalism

Our findings provide a political complement to cognitive explanations for incrementalism in data-driven organizations. A substantial literature, building on the Carnegie tradition (Baldwin & Clark, 1994; March, 1991) attributes incrementalism to methodological and cognitive factors. Analytical techniques such as A/B testing require clean operationalization and historical baselines, making radical innovations difficult to test while incremental changes can be easily measured (Allen & McDonald, 2025; Felin et al., 2020; Wu et al., 2020). These cognitive constraints are real and important. However, this explanation cannot fully account for the incrementalism we observed. iLabor possessed the technical capacity to test Project MATCH's core premise, the structured taxonomy would improve matching quality, and launched an MVP designed to do so. Yet the organization retreated from radical reform not because the experiment was impossible to design, but because it triggered political dynamics that fractured the coalition.

We identify a specific political mechanism: epistemic fragility creates a tradeoff between experimental insight and coalition stability. When project leads pursue ambitious structural reforms in data-driven organizations, they face pressure to demonstrate that the initiative is evidence-

based. But, as we have shown, experiments that address central questions are inherently risky: they attract attention from multiple stakeholders and may produce ambiguous results due to complex interdependencies or long-time horizons of adoption. This can quickly destabilize the authority of the experimental ideology sustaining the coalition. This political dynamic operates alongside cognitive constraints. Even in domains where radical innovation could theoretically be tested, organizations may avoid doing so because the political risks outweigh the epistemic benefits. The result is incrementalism driven not by what organizations can learn through experimentation, but by what they can sustain politically while appealing to it. As scholars in the Carnegie tradition have recently emphasized, organizational cognition and politics are deeply intertwined (Levinthal & Pham, 2024; Levinthal & Rerup, 2021; Rerup & Zbaracki, 2021). We illustrate one manifestation: the very practices meant to enable evidence-based structural reform can trigger political dynamics that make such reform more difficult.

Scope Conditions and Future Directions

Our argument rests on features particularly salient in platform contexts. iLabor's fluid structure, weak hierarchies and high role mobility meant project leads could not rely on occupational jurisdiction or managerial authority to secure buy-in. Where formal authority or professional jurisdiction is stronger, appeals to experimental ideology may matter less. Where experimentation faces fewer constraints (e.g., stand-alone digital products), scrutiny may not unravel the experimental ideal as sharply. In these ways, iLabor is an extreme case. But precisely for that reason, it highlights a core tension in data-driven decision making: scrutiny is central to the scientific method, yet can erode the political authority of data, making "objective evidence" hard to stabilize as neutral authority in goal conflicts. While organizational mechanisms may temper these

dynamics, epistemic fragility remains a risk wherever data-driven decision making becomes a guiding principle.

Several productive research avenues emerge. First, when and how does experimental ideology become re-enchanted after episodes of epistemic fragility? Project MATCH's difficulties did not trigger broad disillusionment with data-driven decision making across iLabor. The ideology persisted in other projects and contexts. When an ideology becomes constitutive of organizational identity (Turco, 2012), localized failures may be quarantined as exceptions rather than processed as disconfirming evidence. Ideology may also be protected through attribution processes that blame individuals or circumstances rather than the ideology itself. Examining such protective dynamics would advance theory on organizational ideology durability despite empirical setbacks.

Second, can organizations escape this dilemma through governance innovations that apply to experimentation and other scientific management techniques? Centralized expert boards, communication strategies, or structural solutions may enable organizations to tolerate ambiguous results on high-stakes questions while maintaining coalitional stability, whether for experiments, machine learning models, or algorithmic systems. Assessing whether epistemic fragility represents a general phenomenon would address concerns about algorithmic governance while advancing theory. Third, the temporal authority invoked by deferral and promise warrants further development. Under what conditions can these strategies sustain coalitions indefinitely versus requiring eventual resolution? How do temporal horizons affect legitimacy? What happens when multiple competing promises circulate? A fuller theory of temporal authority would enrich our understanding of organizational politics beyond data-driven contexts.

More broadly, our claim that experimental ideology is most powerful when practices of experimentation remain unexamined poses fundamental questions for evidence-based governance.

As data-driven decision making spreads across organizations and public institutions, understanding how scientific authority works politically, not just epistemically, becomes critical. The challenge is not simply producing better experiments or data, but developing political and epistemic infrastructures that sustain collective action even when the methods that legitimate it cannot meet their own ideals of transparency and scrutiny indefinitely. Stabilizing common, rigorous standards of evidence within political negotiation is central to the contemporary 'crisis of expertise' (Eyal, 2019) and deserves sustained attention from scholars of organizations and expertise.

REFERENCES

Allen, R. T., & McDonald, R. M. (2025). Methodological Pluralism and Innovation in Data-Driven Organizations. *Administrative Science Quarterly*, *70*(2), 403–443.

Anthony, C. (2021). When Knowledge Work and Analytical Technologies Collide: The Practices and Consequences of Black Boxing Algorithmic Technologies. *Administrative Science Quarterly*, *66*(4), 1173–1212. https://doi.org/10.1177/00018392211016755

Baldwin, C. Y., & Clark, K. B. (1994). Capital-budgeting systems and capabilities investments in US companies after the Second World War. *Business History Review*, *68*(1), 73–109.

Barley, S. R., & Kunda, G. (2001). Bringing Work Back In. *Organization Science*, *12*(1), 76–95. https://doi.org/10.1287/orsc.12.1.76.10122

Barnard, C. (1938). *The Functions of the Executive*. Harvard University Press.

Bechky, B. A. (2003). Object lessons: Workplace artifacts as representations of occupational jurisdiction. *American Journal of Sociology*, *109*(3), 720–752.

Bechky, B. A. (2020). Evaluative spillovers from technological change: The effects of "DNA envy" on occupational practices in forensic science. *Administrative Science Quarterly*, *65*(3), 606–643.

Bower, J. L. (1972). *Managing the resource allocation process: A study of corporate planning and investment*. Irwin Homewood, IL.

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in numbers: How does data-driven decisionmaking affect firm performance? *Available at SSRN: Https://Ssrn.Com/Abstract=1819486 or Http://Dx.Doi.Org/10.2139/Ssrn.1819486*.

Brynjolfsson, E., Jin, W., & McElheran, K. (2021). The power of prediction: Predictive analytics, workplace complements, and business performance. *Business Economics*, *56*(4), 217–239. https://doi.org/10.1057/s11369-021-00224-5

Brynjolfsson, E., & McElheran, K. (2019). Data in action: Data-driven decision making and predictive analytics in US manufacturing. *Rotman School of Management Working Paper*, *3422397*.

Camuffo, A., Cordova, A., Gambardella, A., & Spina, C. (2020). A Scientific Approach to Entrepreneurial Decision Making: Evidence from a Randomized Control Trial. *Management Science*, *66*(2), 564–586. https://doi.org/10.1287/mnsc.2018.3249

Carruthers, B. G., & Espeland, W. N. (1991). Accounting for Rationality: Double-Entry Bookkeeping and the Rhetoric of Economic Rationality. *American Journal of Sociology*, *97*(1), 31–69. https://doi.org/10.1086/229739

Christin, A. (2020). *Metrics at Work—Journalism and the Contested Meaning of Algorithms*. Princeton University Press.

Cyert, R., & March, J. (1963). *A Behavioral theory of the firm*. Martino Publishing.

Daston, L., & Galison, P. (1992). The image of objectivity. *Representations*, *40*, 81–128.

DiBenigno, J. (2018). Anchored Personalization in Managing Goal Conflict between Professional Groups: The Case of U.S. Army Mental Health Care. *Administrative Science Quarterly*, *63*(3), 526–569. https://doi.org/10.1177/0001839217714024

Dougherty, D. (1992). Interpretive Barriers to Successful Product Innovation in Large Firms. *Organization Science*, *3*(2), 179–202. https://doi.org/10.1287/orsc.3.2.179

Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. *Communication Monographs*, *51*(3), 227–242. https://doi.org/10.1080/03637758409390197

Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology*, *24*(1), 313–343.

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology/Archives Européennes de Sociologie*, *49*(3), 401–436.

Ethiraj, S. K., & Levinthal, D. (2009). Hoping for A to Z While Rewarding Only A: Complex Organizations and Multiple Goals. *Organization Science*, *20*(1), 4–21. https://doi.org/10.1287/orsc.1080.0358

Eyal, G. (2019). *The crisis of expertise*. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=Wwu5DwAAQBAJ&oi=fnd&pg=PP2&dq=eyal+crisis+of+expertise&ots=ge3QEnOmM5&sig=BF1uyDNpO8FYWlji2d3c2IFvMmE

Feldman, M. S., & March, J. G. (1981). Information in organizations as signal and symbol. *Administrative Science Quarterly*, 171–186.

Feldman, S. (2004). The Culture of Objectivity: Quantification, Uncertainty, and the Evaluation of Risk at NASA. *Human Relations*, *57*(6), 691–718. https://doi.org/10.1177/0018726704044952

Felin, T., Gambardella, A., Stern, S., & Zenger, T. (2020). Lean startup and the business model: Experimentation revisited. *Long Range Planning*, *53*(4), 101889.

Fourcade, M., & Healy, K. (2024). *The Ordinal Society*. Harvard University Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine Transaction.

Greve, H. R. (2008). A Behavioral Theory of Firm Growth: Sequential Attention to Size and Performance Goals. *Academy of Management Journal*, *51*(3), 476–494. https://doi.org/10.5465/amj.2008.32625975

Guala, F. (2005). *The methodology of experimental economics*. Cambridge University Press.

Hall, T. A., & Hasan, S. (2020). The politics of experimentation. *Available at SSRN 3571296*.

Huising, R. (2014). The Erosion of Expert Control Through Censure Episodes. *Organization Science*, *25*(6), 1633–1661. https://doi.org/10.1287/orsc.2014.0902

Kaplan, S. (2008). Framing contests: Strategy making under uncertainty. *Organization Science*, *19*(5), 729–752.

Karp, R. (2025). *Crossing the Design-Use Divide: How Process Manipulation Shapes the Design and Use of AI*. https://www.hbs.edu/ris/Publication%20Files/25-034_18b7f0b4-fdb6-4225-98ad-2fa7d4adaf42.pdf

Kellogg, K. C. (2011). Hot Lights and Cold Steel: Cultural and Political Toolkits for Practice Change in Surgery. *Organization Science*, *22*(2), 482–502. https://doi.org/10.1287/orsc.1100.0539

Kim, H., Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. (2024). Decision authority and the returns to algorithms. *Strategic Management Journal*, *45*(4), 619–648. https://doi.org/10.1002/smj.3569

Kohavi, R., Crook, T., Longbotham, R., Frasca, B., Henne, R., Ferres, J. L., & Melamed, T. (2009). Online experimentation at Microsoft. *Data Mining Case Studies*, *11*(2009), 39.

Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press.

Koning, R., Hasan, S., & Chatterji, A. (2022). Experimentation and start-up performance: Evidence from A/B testing. *Management Science*, *68*(9), 6434–6453.

Koppman, S., Bechky, B. A., & Cohen, A. C. (2022). Overcoming Conflict Between Symmetric Occupations: How "Creatives" and "Suits" Use Gender Ordering in Advertising. *Academy of Management Journal*, *65*(5), 1623–1651. https://doi.org/10.5465/amj.2020.0806

Latour, B. (1987). *Science in Action—How to Follow Scientists and Engineers Through Society*. Harvard Business Press.

Leifer, E. M. (1991). *Actors as observers: A theory of skill in social relationships*. Garland.

Leonardi, P. M., & Treem, J. W. (2020). Behavioral visibility: A new paradigm for organization studies in the age of digitization, digitalization, and datafication. *Organization Studies*, *41*(12), 1601–1625.

Levina, N., & Orlikowski, W. J. (2009). Understanding shifting power relations within and across organizations: A critical genre analysis. *Academy of Management Journal*, *52*(4), 672–703.

Levina, N., & Vaast, E. (2005). The emergence of boundary spanning competence in practice: Implications for implementation and use of information systems. *MIS Quarterly*, 335–363.

Levinthal, D. A., & Pham, D. N. (2024). Bringing Politics Back In: The Role of Power and Coalitions in Organizational Adaptation. *Organization Science*, *35*(5), 1704–1720. https://doi.org/10.1287/orsc.2022.16995

Levinthal, D. A., & Rerup, C. (2021). The Plural of Goal: Learning in a World of Ambiguity. *Organization Science*, *32*(3), 527–543. https://doi.org/10.1287/orsc.2020.1383

Luca, M., & Bazerman, M. H. (2021). *The Power of Experiments: Decision Making in a Data-Driven World*. MIT Press.

Luca, M., & Edmondson, A. C. (2024). Where Data-Driven Decision-Making Can Go Wrong. *Harvard Business Review*, *102*, 80–89.

March, J. G. (1962). The Business Firm as a Political Coalition. *The Journal of Politics*, *24*(4), 662–678. https://doi.org/10.1017/S0022381600016169

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, *2*(1), 71–87.

March, J. G., & Simon, H. A. (1958). *Organizations*. JohnWiley & Sons.

Marres, N., & Stark, D. (2020). Put to the test: For a new sociology of testing. *The British Journal of Sociology*, *71*(3), 423–443.

Mazmanian, M., & Beckman, C. M. (2018). "Making" Your Numbers: Engendering Organizational Control Through a Ritual of Quantification. *Organization Science*, *29*(3), 357–379. https://doi.org/10.1287/orsc.2017.1185

Mennicken, A., & Espeland, W. N. (2019). What's new with numbers? Sociological approaches to the study of quantification. *Annual Review of Sociology*, *45*(1), 223–245.

Mithani, M. A., & O'Brien, J. P. (2021). So What Exactly Is a "Coalition" Within an Organization? A Review and Organizing Framework. *Journal of Management*, *47*(1), 171–206. https://doi.org/10.1177/0149206320950433

Nigam, A., Huising, R., & Golden, B. (2016). Explaining the Selection of Routines for Change during Organizational Search. *Administrative Science Quarterly*, *61*(4), 551–583. https://doi.org/10.1177/0001839216653712

O'Neill, C. (2016). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy; Crown*. Penguin Random House LLC: New York, NY, USA.

Pinch, T. (1993). "Testing-One, Two, Three... Testing!": Toward a Sociology of Testing. *Science, Technology, & Human Values*, *18*(1), 25–41.

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Rahman, H. A., Weiss, T., & Karunakaran, A. (2023). The experimental hand: How platform-based experimentation reconfigures worker autonomy. *Academy of Management Journal*, *66*(6), 1803–1830.

Ranganathan, A., & Benson, A. (2020). A Numbers Game: Quantification of Work, Auto-Gamification, and Worker Productivity. *American Sociological Review*, *85*(4), 573–609. https://doi.org/10.1177/0003122420936665

Ransbotham, S., Kiron, D., & Prentice, P. K. (2016). Beyond the Hype: The Hard Work Behind Analytics Success. *MIT Sloan Management Review*. https://sloanreview.mit.edu/projects/the-hard-work-behind-data-analytics-strategy/

Reinecke, J., & Donaghey, J. (2025). From constructive ambiguity to escalating commitment: The evolution of the Bangladesh accord as a transnational institution for collective action. *Administrative Science Quarterly*, 00018392251331027.

Rerup, C., & Zbaracki, M. J. (2021). The Politics of Learning from Rare Events. *Organization Science*, *32*(6), 1391–1414. https://doi.org/10.1287/orsc.2020.1424

Sitkin, S. B. (1992). Learning through failure: The strategy of small losses. *Research in Organizational Behavior*, *14*, 231–266.

Stice-Lusvardi, R., Hinds, P. J., & Valentine, M. (2024). Legitimating Illegitimate Practices: How Data Analysts Compromised Their Standards to Promote Quantification. *Organization Science*, *35*(2), 432–452. https://doi.org/10.1287/orsc.2022.1655

Strauss, A., & Corbin, J. (2007). *Basics of qualitative research techniques* (3rd ed.). SAGE Publications, Inc.

Thomke, S. H. (2001). Enlightened experimentation. The new imperative for innovation. *Harvard Business Review*, *79*(2), 66–75.

Thomke, S. H. (2020). *Experimentation works: The surprising power of business experiments*. Harvard Business Press.

Truelove, E., & Kellogg, K. C. (2016). The Radical Flank Effect and Cross-occupational Collaboration for Technology Development during a Power Shift. *Administrative Science Quarterly*, *61*(4), 662–701. https://doi.org/10.1177/0001839216647679

Turco, C. (2012). Difficult Decoupling: Employee Resistance to the Commercialization of Personal Settings. *American Journal of Sociology*, *118*(2), 380–419. https://doi.org/10.1086/666505

Van Den Bulte, C., & Moenaert, R. K. (1998). The Effects of R&D Team Co-location on Communication Patterns among R&D, Marketing, and Manufacturing. *Management Science*, *44*(11-part-2), S1–S18. https://doi.org/10.1287/mnsc.44.11.S1

Van Maanen, J. (2011). *Tales of the field: On writing ethnography*. University of Chicago Press.

Wu, L., Hitt, L., & Lou, B. (2020). Data analytics, innovation, and firm productivity. *Management Science*, *66*(5), 2017–2039.

Zbaracki, M. J. (1998). The Rhetoric and Reality of Total Quality Management. *Administrative Science Quarterly*, *43*(3), 602–636. https://doi.org/10.2307/2393677

TABLES AND FIGURES

Table 1. Occupational Groups working on Project MATCH

Functional Group	Count
Product	
Product Managers	7
UX Research	4
UX Design	9
Engineering	
Engineering Managers	10
Front-End Engineers	27
Back-End Engineers	16
Data Scientists	6
Data Analytics	1
Operations	2
Product Marketing	3
Total	**85**

Table 2. Overview of Workstreams and Associated Occupational Groups

Workstream	Focus	Occupational Groups
1	Redesign of Freelancer Profile, Freelancer Search UI, Client Job Post, Job Post UI	Product, Design, & Front-End Engineering
2	Creation of a Taxonomy Management System	Product & Back-End Engineering
3	Recalibrating Algorithms for Freelancer Search	Product, Data Science & Back-End Engineering
4	Recalibrating Algorithms for Job Search	Product, Data Science & Back-End Engineering
5	Improving accuracy of Supply-Demand Dashboards and setting new Governance Rules	Product, Operations, Data Analytics, & Data Science
6	Introducing product changes to the market and conducting User Research	Product, Customer Research & Product Marketing

Table 3. Overview of Data Sources and their Use in the Analysis

Source	Workstream	Use in the Analysis
Observations: 8 months onsite observation complemented by remote fieldwork	MATCH: 197 design-related meetings (136 hrs), including project, unit, and supervisory meetings, design sprints, internal demos, team debriefs, cross-functional meetings, roadmap reviews, as well as sessions with external consultants	Core data source about how work was organized and carried out across workstreams. Familiarization with the language, data, and metrics used to construct shared understanding. Insight into real-time collaboration, project demands, decision making, conflict resolution, including the strategies used to resolve goal conflicts.
	iLabor: new employee introduction days, all staff gatherings, manager trainings, a women's summit, offsite events and strategic roadmapping sessions, product meetings and retrospectives, and a one-week annual engineering meetup	Familiarization with iLabor's organizational culture, values, and strategic priorities. Offered insight into how leadership framed initiatives, how norms and expectations were communicated, and how organizational members engaged across hierarchy and roles. Enriched our understanding of Project MATCH by situating it within the broader organizational context.
Interviews: 107 interviews (86 hrs) with 54 members	MATCH: interviews with product managers (26), UX designers (14), engineers (front- and back-end; 30), data scientists (6), data analysts (4), product marketing specialists (3), and customer researchers (10)	Deepened understanding of how role-based values and cognitive frames shaped the work across workstreams. Helped uncover idealized visions for the platform's future, perceived tradeoffs in Project MATCH, and backstage struggles. Provided insight into how members narrated the authority and limits of data and experimentation, and interviewees perceived their role in decision making, resolving internal tensions, and aligning interests across roles and workstreams.
	iLabor: interviews with the leadership team (14)	Highlighted how senior leaders framed organizational goals and the role of data in product design. Revealed how leaders soguht to align the work in Project MATCH with high-level objectives.
Work materials, online and archival data: 694 project documents and 87 external communications	MATCH: meeting screenshots, internal presentations and overview documents, meeting notes and reports, product specifications, Jira tickets, email threads, Slack messages, design mock-ups, work calendars, and engineering documentation (694 documents)	Offered fine-grained detail on hoe Project MATCH was envisioned, scoped, debated, and implemented. Familiarization with workplace communication infrastructure. Identification of design inconsistencies across workstreams. Triangulation of the evidence derived from interviews and observations.
	iLabor: 87 communications sent to iLabor users between 2017 and 2020, including onboarding messages, product announcements, and update notifications	Helped understand how product changes were translated into user-facing narratives and promises, revealing how Project MATCH was positioned as a major platform improvement.

Table 4. Strategies of invoking the ideology of experimentation

Strategy	Definition	Purpose
Assert experimental authority	Leverage the prestige of experimentation to legitimize a preferred position or discount alternatives from non-experimental evidence—even when the cited evidence does not meet experimental standards—in order to close debate for the moment.	Using authority to close debate and secure a decision over the best course of action
Defer pending evidence	Postpone a contested decision by citing the absence of definitive experimental evidence and committing to future testing to maintain coalition buy-in	Using ambiguity to suspend currently irreconcilable goal conflicts and maintain buy-in
Promise future resolution	Proceed with a provisional decision now while committing that iterative, experiment-driven development will ultimately deliver the ideal solution, rendering today's conflicts irrelevant.	Using ambiguity to bracket currently irreconcilable goal conflicts to maintain project momentum

Table 5. Design Issues where Project Leads Deployed Strategies of Experimental Authority, Deferral, and Promise

Issue		Strategy	Before MVP	After MVP	Illustrative examples
1	***Do you allow for users to set the categories of the taxonomy?***				
Back-End	No, user-generated content is poorly structured and will produce inconsisten- cies with the decided system.	Defer pending evidence	✓		In this discussion on freelancer search, a data scientist worried about inconsistencies that would arise from mixing user-generated and system-generated terms in the taxonomy. If the taxonomy contained both a controlled and an uncontrolled vocabulary, it would become difficult to determine what the system should optimize matches on. As data scientist Yiorgos put it: "My point is, you need to decide at some point, how to design a metric around what a good profile looks like" [20170608-O-PROD-EXT]. External consultant Dustin suggested to postpone the issue. He argued that it was an empirical question which terms could reliably indicate quality, and that the team should remain agnostic until they had data. He suggested that both a centrally controlled taxonomy and a user-generated taxonomy might be "equal or irrelevant" for this purpose and that "when you get to a point where you actually can't distinguish [what captures applicants' quality], you might as well use the whole diversity [of terms] in order to optimise for things other than the direct relevance." In his view, "the reasons that favour one ordering [of terms] over another are dominated by the ecosystem," and these could only be revealed through careful data collection and analysis—at some later stage. Product manager Zane then showed his agreement, before moving on to the next discussion topic [20170608-O-PROD-EXT]. In this case, the appeal to defer pending evidence succeeded.
Market governance	Yes, we will use a combination of centralized planning and machine learning to infer the taxonomy from user behavior.				
2	***Are freelancers allowed to retain their general profile, for how long (forced adaptation)?***				
Front-End	Yes. Retain the general profile indefinitely to maximize the ways in which freelancers can compete	Defer pending evidence	✓		In this example, back-end engineer Pavel suggests that there are logical inconsistencies between the general and the specialized profile. If clients can see both, they will not know how to assess this inconsistency. UX designer Heath translates the issue into an empirical question and then defers: "Yeah, I think that's definitely something to watch out for during testing: what the clients think a [general] profile is, what they think it means. For sure, [...] once we explained it correctly, and we have the right label on there, most of these questions I think will go away." Pavel accepts this argument, stating: "Yeah, yeah, everything is logical," adding that "The main key here is whether it is mandatory to have a specialized profile to apply to a specialized service job or not. If it is not mandatory then there are no problems." [20170615-O-ENG-CFD]
Back-End	No, get rid of general profile to avoid inconsistent data tables				
Market governance	No. The general profile violates the core idea behind project MATCH because it creates competition outside the categories of the taxonomy.				

Issue		Strategy	Before MVP	After MVP	Illustrative examples
3	***How granular should the work categories of the taxonomy be?***				
Front-End	Not granular because we do not want to overwhelm users.	Defer pending evidence	✓	✗	Product managers initially appealed to deferral by framing this conflict as an empirical question to be resolved through testing. As product manager Zane explained, "the level of specificity [for any given category] is a function of the volume of the work on the platform and the flexibility of the workers in the real world" [20170616-O-CFT]. After the MVP launch, however, engineers resisted this line of reasoning. They argued that testing for the appropriate level of granularity conflicted with the need to keep categories up to date. By the time enough users had adopted a set of categories to evaluate its granularity, those categories would already be outdated and of little use to freelancers and clients. As engineer Karlo put it: "That's the biggest problem of it. We want to do things quickly, want to adjust [to market changes] quickly. But the problem is [that] the major metric that we want to look at [that is of appropriate granularity] can [only] be seen really after a longer period of time" [20171122-I-Karlo]. In this context, testing was no longer viewed as a panacea for all disagreements.
Back-End	Not granular because highly differentiated system is difficult to manage.				
Market Governance	Very granular means better market legibility, which is good.				
4	***Do we let freelancers curate how their data inputs are presented to clients in search?***				
Front-End	Yes. It is important that freelancers retain control over the criteria they can be matched on.	Defer pending evidence	✓		In this example, the design team suggested that they should signal to freelancers how the information they input might appear differently to clients. UX designer Audrey asked: "Would that be a way to convey the [discrepancy] to our freelancer: There's search tile and there are actually two versions. [In one] they can see if a client searches for a term, they will see me as represented in this [first] search tile." Other members of the team reinforced the concern. Product manager Celine recalled: "Yeah, I remember hearing feedback from some of the users that they were concerned that the lingo they were using in their profile isn't going to match with the client searches. So I agree: I think we need to do the fancy footwork, and give them a visual reassurance that they will be found either way." Not wanting to press the issue, product manager Zane deferred pending evidence, suggesting they experiment with the degree to which freelancers could see the difference between what they wrote and what clients saw in search results: "Maybe we do like a sampling of terms under which they will appear?" UX designer Shruti agreed: "Yeah, I think we could review [in our next test] how it looks in search to show all the different keywords that the client might be searching for" [20170623-O-PROD -DESIGN]. In this case, Zane's appeal to deferral successfully postponed the issue.
Market governance	No, we need to hide this from freelancers to optimize matches for clients and avoid gaming				

Issue		Strategy	Before MVP	After MVP	Illustrative examples
5	***Who decides when to reassign freelancers from one category to another?***				
Front-End	Freelancers do: they can switch and choose categories as they want because they know best	Defer pending evidence	✓		In this design thinking meeting, designers, customer researchers, and product managers discussed persona exercises where they created job posts and specialized profiles for different hypothetical professionals. At one point, UX designer Audrey raised concerns about how freelancers would be assigned to categories. She argued that if control were imposed from above, freelancers would need to be strategic about how they categorized themselves, but in practice, most simply tried to represent their work history accurately: "Embedded in what Heath said, is that Freelancers will be strategic about spreading out their work history in a way that will get them more jobs. I think you're assuming that they will understand that they have to be strategic here, [i.e. make tradeoffs like] either double down on email support, or try and promote themselves as a chat support, versus take a more straightforward approach of just thinking: I'm genuinely trying to decide what was primary about this job, so I can categorise it correctly. Right. I don't know if freelancers will uniformly make that leap to thinking about this as marketing versus thinking about this as accurate representation." [20170707-O-CR- Design Thinking Session]. UX designer Heath appealed deferred pending evidence, suggesting that this distinction would become clearer only once the product was tested in the market: "At first maybe not, I think, eventually, that'll become more clear [as we do more testing]. But it's not like we're trying to incentivize them to game the system." [20170707-O-CR-Design Thinking Session]. In this case, Hayden's deferral successfully postponed the issue.
Market governance	We do: When there are shortages of freelancers we can redraw the boundaries of the categories				
6	***How hierarchical should the taxonomy be? How exclusive should the relationship between skills, specializations, and attributes be?***				
Front-End	Not hierarchical at all: to allow maximum autonomy the system should be unstructured.	Promise future resolution & Assert experimental authority	✓		When back-end engineers raised questions about the adoption of a hierarchical taxonomy system, worrying that users would not want to learn how categories and subcategories related to each other, product manager Zane promised future resolution to settle the issue. He suggested issues would be resolved with the hierarchy being intuitive in the final version: "We would construct hierarchies [of concepts] that meet people where they're trying to meet us, not force them to learn specifically how we arranged everything" [20170914-O-OFFSITE].
Back-End	Not hierarchical: We don't know exactly how the taxonomy will develop, so we should choose a less structured system.				

Issue		Strategy	Before MVP	After MVP	Illustrative examples
Market governance	Very hierarchical but with enough flexibility to allow users to arrive at the same descriptions from multiple directions (graph structure).				Still skeptical, the head of product, Harper pressed further: "How would you validate that that is working?" Zane began outlining experiments to test the intuitive validity of the taxonomy. Head of engineering, Hui chimed in, "I suspect we could like keep this conversation going for another six hours on like how can we make this more awesome. But it also sounds like we are not sure... And there is an experimentation approach we can rely on. . . Do we have a sense of what would be the definition of success [and] what do we have in the experimental pipeline? Zane indeed appealed to experimental authority in outlining the experimental phase of h2 in the product, specifically in the search funnel, such as start to end performance of the new search funnel as well as "moments inside of this funnel as they tell you different things about the performance of the product, and the performance of the business." He continued, "what we're looking to do is to improve that ratio of impression to click to invite. And the more invites we get out of this system with less wasted effort on the client side of the experience, the better we think we're doing in the match" [20170914-O-OFFSITE]. In this context, asserting experimental authority and promising future resolution helped settle the issue.
7	***Can freelancers apply to jobs outside their specialized profile?***				
Front-End	Yes, it otherwise constrains the user experience	Promise future resolution	✓		In this example, UX researcher Cleo pushed back against plans to limit freelancers to jobs in their chosen specialization. She argued: "[One critical issue is] whether we can have a system that will encourage or incentivize our freelancers and clients to play along. […] If all the information is there and we cannot get people to fill out the information, then that is not really going to work." If users did not complete their new profiles, she explained, the matches would not improve, and freelancers would resent any restrictions on the classic ways of searching for jobs [20170914-O-PROD-ENG-OFFSITE]. Product manager Zane promised future resolution for this concern. He acknowledged that users would be angry but framed this as a temporary, "in-between state" loss: "Yeah,
Back-End	No, it is inconsistent with the logic of the search algorithm				

Issue		Strategy	Before MVP	After MVP	Illustrative examples
Market Governance	No. The whole point of Project MATCH is specialization.				I think this [anger] is something I would expect to experience during the transitional period where profiles go from optional to mandatory. There [will be] a lot of friction and a lot of unhappiness from freelancers who, you know, didn't heed our warnings, or just didn't get around to [filling out their profiles]. And all of a sudden they can't apply to jobs; they can't get invitations; that's going to be very upsetting." He then suggested that this frustration would ultimately create the incentive for freelancers to complete their profiles, thereby improving the system: "Frankly, even that, I think, is a beneficial outcome of imposing the system on the market." [20170914-O-PROD-ENG-OFFSITE] In this case, Zane's appeal to idealization succeeded.
8	***When to change the taxonomy?***				
Back-End	Not as frequently, so that we can ensure greater system stability.	Promise future resolution	✓	✗	Before the MVP, engineers showed that frequent changes to the taxonomy could introduce inconsistencies into a user's job-post workflow. As engineer Oleh explained: "when the user saves [the first] page to create a job post, he wants to see the subject line on the next page created with the appropriate specialisations and skills." Because the user's initial choice needed to recur in many different contexts, the taxonomy elements had to display "strong consistency" [20170711-O-ENG-back-end]. In response, product managers appealed to promise future resolution, suggesting that the final product would naturally ensure consistent workflows despite frequent changes: "The job post form itself will eventually [be] consistent, right? […] Eventual consistency is just a pattern we have to adopt on all of our systems anyway. So, we'll just work around the current consistency issues." From this perspective, inconsistencies in the user interface were seen as nothing more than "in-between state losses" on the path toward experimental perfection [20170711-O-ENG- back-end]. After the MVP, however, engineers no longer believed that iterative development could deliver the desired consistency. They argued that frequent changes created too many problems to be resolved through testing. As engineer Karlo put it: "Like we're constantly testing for [the consistency of] job post, it always says we do [but] it can be affected by so many different things" [20171122-I-Karlo]. In this case, appeals to promise future resolution first succeeded but later failed.
Market governance	Frequently, so that the taxonomy system is always in line with the use of the categories by users.				
9	***Do you constrain freelancers' ability to choose their specializations based on other information on their profile?***				

Issue		Strategy	Before MVP	After MVP	Illustrative examples
Front-End	No, give users autonomy to redefine themselves	Assert experimental authority & Defer pending evidence		✗	Front-end wanted no hard limits; back-end favored prefills or restrictions. Head of engineering Hui pressed: "I was surprised . . . there is no way . . . you've ever helped me [a user] prefill what I already have put into the box." A UX designer replied: "We knew that if we prefilled, people would just skip." A product manager first appealed to experimental authority, "we definitely heard it from user testing," and then deferred pending evidence, "it's good to go out with this and then see the kind of titles we're getting out." Hui countered this with suggesting a measurable alternative (detect no-edit prefills): "Hey, we won't let you do this because you didn't change anything." In this case, asserting experimental authority and defer pending evidence failed [20171110-O-ENG].
Back-End	No, users should actively migrate data into the new taxonomy to ensure consistency				
Market governance	Yes, structure is good. Give some control, but restrict actions based on preexisting data				
10	***How many profiles do you allow freelancers to have?***				
Front-End	As many as they want. If matching doesn't improve, they would be angry at arbitrary limitations to their ability to compete	Promise future resolution		✗	In a "Demo & Donuts" session, product managers defended their compromise with front-end not to limit the number of freelancer profiles initially, even though this advantaged multi-skilled freelancers and disadvantaged others. The inequity stemmed from the taxonomy: the number of profiles depended on which categories were recognized, and these categories represented an arbitrary creation at this stage. In a chat, an engineer raised the issue. Samantha read it out to the presenters. Samantha: "So as far as I know, we have around 120 categories of work, and then many more subcategories. We're going to create specialised profiles for all of these categories. How do you make the selection and how long it will take to create specialised profiles for every category out there?" Celine initially tried to sidestepp the deeper issue by focusing on the second part of the question about sequencing of rollout. But the engineer came back to the core issue that the decisions about categories affected how many profiles freelancers could compete with: Samantha: "And so basically, freelancers that have few skills, they basically will have few profiles?" This implictly raised the contentious issue that these freelancers would be disadvantaged relative to those whose skills corresponded to established categories. Celine turned to the strategy of promise to resolve the tension: "Yes. Yeah. And so we've already seen a few examples where people who are creating these mobile development, specialisations actually have other skills like translation or graphic design, and we want them to be able to have
Back-End	Limit to a fixed number to avoid imbalanced competition that derives from arbitrary construction of the category system.				

Issue		Strategy	Before MVP	After MVP	Illustrative examples
Market governance	Not many. We will limit the number to encourage specialization				specialised profiles for that, too. We'll definitely limit it because we want people to be focused a lot of people who tend to claim translation, it's because they're just starting out on ILabor and trying to build up a reputation. But we're hoping that once they can be more focused, they only need maybe two or three specialised profiles." In this response, she suggests that the problem will be resolved in the future becasue they will limit the number of profiles and freelancers won't have use for them anymore anyways. The taxonomy will be well developed enough to capture the relevant market segmentation perfectly. Yet this effort to promise resolution at the future point of full implementation through data-driven design. The audience kept pressing: "How deep are you guys gonna go? I see you have iOS development for, you know, mobile development. Are you going to do the same for marketing, like performance marketing, for product marketing, for marketing. How do you decide?" Defensive and exasperated, one of the product managers on the stage muttered to himself, "Stop asking more questions."
11	***Do you let clients search for freelancers without specialized profile in the relevant category? Which profiles become visible in search? Can clients still see breadth and variety of a given users profile in search?***				
Front-End	Yes: freelancers should be able to advertise the full breadth of their skills	Promise future resolution & Defer pending evidence	✓✗		In this example, product marketeer April objected to a decision to limit freelancer search to show freelancer's specialized profile. She argued that the team had effectively told the community: "Don't worry freelancers, clients will know that you still do multiple things. And that seems to be lost in that. I think freelancers will call us on that, because we're not indicating in any way that they do other stuff" [20171205-O-E2E Review of FL Search]. UX designer Heath attempted to appease her by promising future resolution, reminding her of the larger goals of the project, which would eventually allow freelancers to
Back-End	No: Consistency requires that the two search systems operate separately. We only fall back into the broader search when there is no result from the taxonomy system				

Issue	Strategy	Before MVP	After MVP	Illustrative examples
Market governance	No. Specialized search only surfaces specialized profiles.			express everything they needed within the specialized profile: “I think it is okay. We're just gonna have to figure out if that's a problem or not. I think the whole point of this system is it’s better to say that they specialise in iOS development than say, like ‘hey, I specialise in all this other stuff, if the client is looking for specifically iOS development.” Product manager Celine agreed. UX designer Audrey mentioned this will all get resolved once the full system has been rolled out for all services, but recognizes the issue in this in-between state, and to which April then responded "maybe we file this away as another in-between state loss." Later in the meeting Heath returned to the issue and tried to use deferral pending evidence, calling for more empirical data: “It’s kind of tricky until we see how clients interact with it. Do they pay attention to it? All we already know…” But April cut him off, reiterating her concern: “I just want this to be something successful for freelancers as well. And it is something that when we originally scoped it out, we said [that] we would have [it]” She then nevertheless agreed, stating "it's worth testing, right, and we could do it on the 12th. . . And that's a mindset shift that will come over time, right?" Heath then responded that this problem will indeed be resolved in the future, "Yeah, I think eventually what will happen is, you will only be able to apply to jobs as specialized profiles, which will nullify sort of the general profile, because there'll be no use for it [in search]." [20171205-O-E2E Review of FL Search]. So close before launch, the strategy of defer pending evidence failed but promise future resolution helped to address April's objections for now.

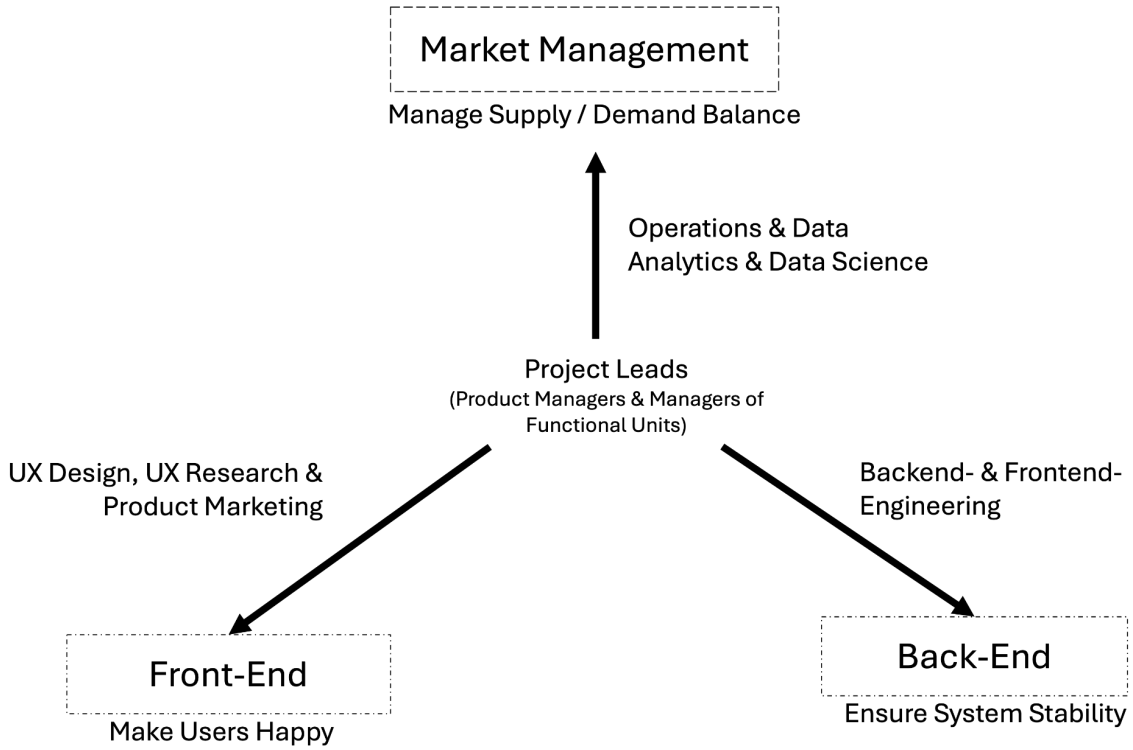

Figure 1 - Value Orientations by Technology Architecture at iLabor

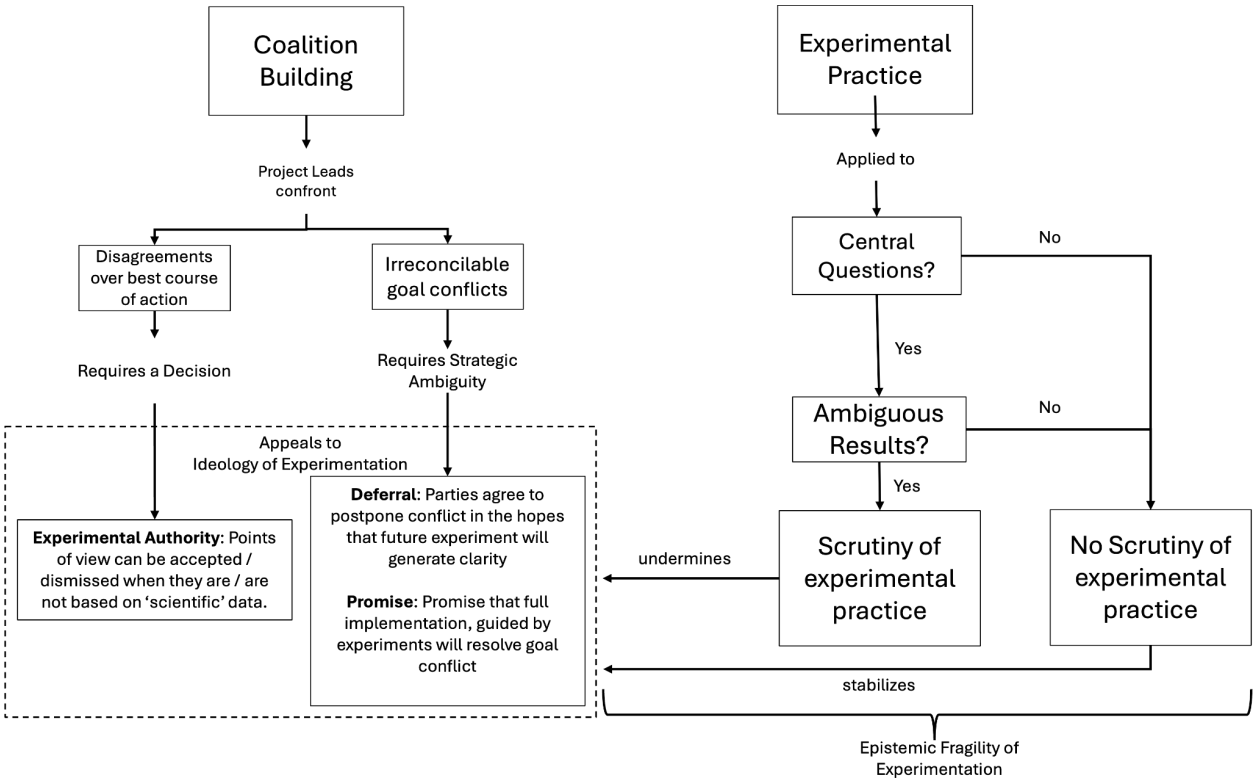

Figure 2 - Theoretical Model: Epistemic Fragility of Experimentation