Scaling up experimental social, behavioral, and economic science

Apr 29, 2021

CSSLab

Authored by: Abdullah Almaatouq, Joshua Becker, Michael S. Bernstein, Robert Botto, Eric T. Bradlow, Ekaterina Damer, Angela Duckworth, Tom Griffiths, Joshua K. Hartshorne, David Lazer, Edith Law, Min Liu, J. Nathan Matias, David Rand, Matthew Salganik, Emma Satlof-Bedrick, Maurice Schweitzer, Hirokazu Shirado, Jordan W. Suchow, Siddharth Suri, Milena Tsvetkova, Duncan J. Watts, Mark E. Whiting and Ming Yin.

Abstract

The standard experimental paradigm in the social, behavioral, and economic sciences is extremely limited. Although recent advances in digital technologies and crowdsourcing services allow individual experiments to be deployed and run faster than in traditional physical labs, a majority of experiments still focus on one-off results that do not generalize easily to real-world contexts or even to other variations of the same experiment. As a result, there exist few universally acknowledged findings, and even those are occasionally overturned by new data. We argue that to achieve replicable, generalizable, scalable and ultimately useful social and behavioral science, a fundamental rethinking of the model of virtual-laboratory style experiments is required. Not only is it possible to design and run experiments that are radically different in scale and scope than was possible in an era of physical labs; this ability allows us to ask fundamentally different types of questions than have been asked historically of lab studies. We argue, however, that taking full advantage of this new and exciting potential will require four major changes to the infrastructure, methodology, and culture of experimental science: (1) significant investments in software design and participant recruitment, (2) innovations in experimental design and analysis of experimental data, (3) adoption of new models of collaboration, and (4) a new understanding of the nature and role of theory in experimental social and behavioral science. We conclude that the path we outline, although ambitious, is well within the power of current technology and has the potential to facilitate a new class of scientific advances in social, behavioral and economic studies.

Table of Contents

Introduction

Over the past half century or so, lab experiments have yielded fundamental insights into decision making (Tversky and Kahneman 1974), social influence (Asch 1951; Milgram 1969), cooperation (Yamagishi 1986), cultural evolution (Kirby, Cornish, and Smith 2008), market dynamics (Plott 1982) and numerous other success stories. More recently, other experimental methods such as field experiments have also become popular in social science (Duflo, Glennerster, and Kremer 2007) and in business (Manzi 2012), where they are often referred to as "A/B tests" (Kohavi, Longbotham, and Walker 2010). Although both types of experiments leverage randomized assignments, lab experiments afford greater control and are therefore often preferred for theory testing and development. However, they also exhibit at least three serious deficiencies with respect to the generalizability of their findings, consistent challenges to both their external and ecological validity.

First, the necessity of recruiting participants who can physically show up at the investigator's lab has historically limited participation in lab experiments to university students — typically undergraduates — and occasionally members of the local community. Experimental findings are therefore heavily biased toward WEIRD (Western, Educated, Industrialized, Rich, and Democratic) societies (Henrich, Heine, and Norenzayan 2010) and are rarely representative of even those populations. Though the findings are often assumed to generalize to other populations, this is not always — and perhaps not even often — the case (Henrich et al. 2001, 2005; Nisbett 2004). The cost and logistical complexity of running typical lab experiments hamper efforts to systematically identify conditions under which results from homogenous and/or non-representative groups can be expected to generalize.

Second, while lab experiments are more flexible than A/B testing, they are still severely constrained. Figure 1 shows the design space of SBE experiments collapsed onto three conceptual dimensions: size (of the population to be studied); duration (over which the experiment runs); and complexity (of the interactions involved). Many problems of interest to the human sciences--for example life-course outcomes, political polarization, democratic decision making, and economic growth and security —involve large, diverse populations of people interacting in complex ways over long periods of time: weeks, months, or even years. Traditional lab experiments, in contrast, are limited to small groups of people, often single individuals, interacting in simplistic ways (responding to a survey, contributing to a public good from some allocated endowment, issuing a judgment on some specified outcome) over time intervals measured in minutes. In light of these limitations, it should come as no surprise that the leap from the actual results of a lab experiment to its real-world implications is necessarily tenuous and speculative (Yarkoni 2019).

Third, the cost and administrative effort required to run a single experiment in a lab places heavy constraints on the number of such experiments that any single researcher or lab can conduct. As a result, individual experiments necessarily make specific choices with respect to parameters that are not the analytical focus of the experiment. For example, a prisoner's dilemma experiment on the effect of reward or punishment will typically focus on a single choice of parameters for the payoff matrix or the game length (Embrey, Fréchette, and Yuksel 2018), while an experiment on decision making under uncertainty will focus on specific parameters of a gamble (Tversky and Kahneman 1974). Over time, different experiments on the same general question will make different choices with regard to these non-focal parameters, implicitly assuming that they have no effect on the results. Almost certainly, however, these degrees of freedom in experiment design do affect the results, thereby contributing to inconsistencies and contradictions across ostensibly comparable studies (Watts 2017; A. M. Almaatouq 2019; Landy et al. 2020). While meta-analyses can resolve some of this disagreement, the absence of systematic data on the differences caused by all such variations in experimental conditions prevents even the most comprehensive meta-analysis from reconstructing the full set of dependencies. Even for relatively simple settings in relatively mature areas of study — say, individual judgment and decision making or two-player games — the state of scientific knowledge is surprisingly fragile with respect to contextual variation (Goroff et al. 2018).

While researchers in SBE sciences have started to take advantage of virtual lab experimentation, often these efforts translate existing research designs from the physical lab to a virtual environment. This approach significantly under-utilizes the potential of virtual labs, which offer an opportunity to recruit orders of magnitude more participants and to execute experiments that could never be done in a

physical laboratory. Our plan to unlock this potential encompasses contributing core infrastructure and architecture and encouraging the research community to take on higher throughput research designs through workshops and collaborative efforts. With these, we aim to respond to the underlying questions about the challenges and roadblocks to scaling up experimental SBE sciences. Further, these new kinds of experiments will offer higher levels of external and ecological validity than experimental approaches in popular use in these fields today.

We believe that these deficiencies can be addressed by supporting the development of a new generation of virtual research labs, operating at a larger scale and with more ambitious approaches to experimentation and collaboration than current efforts. We identify two distinct benefits of resolving these challenges: high throughput virtual labs enable large-scale systematic exploration of experimental spaces — impossible with in-person labs, and with more precision than online field experiments — and high throughput virtual labs enable new kinds of SBE experiments such as new units of analysis and adaptive treatment selection.

The Promise of Virtual Labs

Shortly after the World Wide Web had been invented, researchers began to employ "virtual lab" experiments, in which the traditional model of an experiment conducted in a physical lab is translated into an online environment (Horton, Rand, and Zeckhauser 2011; Musch and Reips 2000; Mason and Suri 2012; Reips 2012; Paolacci, Chandler, and Ipeirotis 2010). Virtual labs are appealing on the grounds that, in principle, they resolve many of the constraints on participant recruitment, research design, and practical administration that arise from the necessity of physically co-locating human participants and the experimenter.

With respect to participant recruitment, crowdsourcing services such as Amazon Mechanical Turk and Prolific now allow researchers to recruit and pay larger and more diverse samples of participants than are available on college campuses or local communities (Mason and Suri 2012). In addition, the speed and cost advantages of crowdsourcing services have allowed researchers to run, in effect, thousands of experiments that systematically cover the parameter space of a given design. For example, this approach was taken in the Choice Prediction Competitions, where human decision-making was studied by automatically generating over 100 pairs of gambles following a predefined algorithm (Erev et al. 2017; Plonsky et al. 2019). Recent work took advantage of the larger sample sizes that can be obtained through virtual labs to scale up this approach, collecting human decisions for over 10,000 pairs of gambles (Bourgin et al. 2019). In other cases, online experiments have attracted large and diverse populations of participants who participate voluntarily out of intrinsic interest. For example, one experiment collected almost forty million decisions from over a million unique participants in over 200 countries (Awad et al. 2018).

With respect to research design, when combined with larger sample sizes, lower costs, and faster turnaround times, the flexibility around time and space afforded by remote participation has enabled researchers to design experiments that would be difficult to run in a physical lab, or even impossible. For example, researchers have succeeded in designing "macro-sociological" experiments in which the unit of analysis is a collective entity such as a market (Salganik, Dodds, and Watts 2006), an organization (Valentine et al. 2017) or a community (Whiting et al. 2017) comprising dozens or even hundreds of individuals. By re-recruiting the same

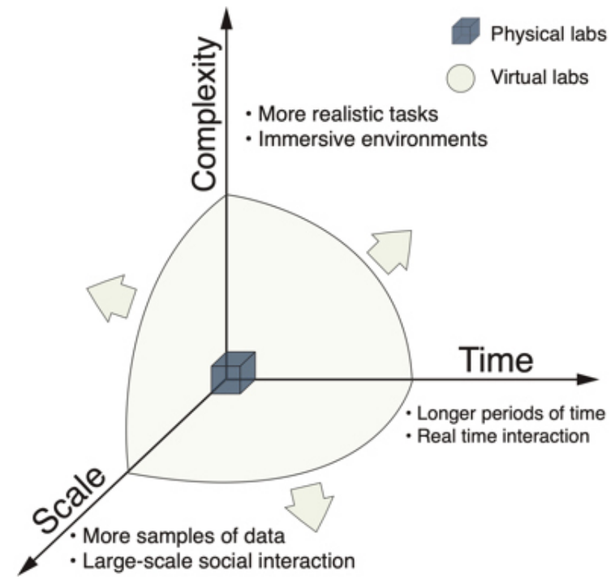

Figure 1. Schematic of the design space of lab experiments. *Reproduced with permission (A. Almaatouq, Becker, et al. 2020)*

participants to return to the "lab" on many separate occasions (i.e. constructing a panel data set) , researchers have also examined how behavior evolves on longer timescales while retaining high-resolution measurements (Mao et al. 2017). Finally, it is possible to recreate complex, immersive, and realistic activities in tightly controlled experimental environments not possible in physical labs (Mao et al. 2016; Whiting et al. 2020). This is, of course, not to say that distinct advantages of in person labs, which are particularly important where control over subjects is important; or where psychological processes invoked by physical proximity are central; or where certain types of physiological interventions of measurements might be necessary.

The capacity for virtual environments to facilitate experiments that are larger, more complex, and more realistic than would be feasible to run in physical labs, and to run these experiments faster and more cheaply, should allow researchers to dramatically expand the accessible design space for experiments, with correspondingly dramatic improvements in the replicability, robustness, and

usefulness of social and behavioral experimental science (see Figure 1). In the remainder of this section, we outline examples of some of the designs that are made possible by the increased sample sizes and experimental control afforded by running experiments online. For conceptual clarity, we have organized these examples in ascending units of analysis, beginning with individuals, then groups, then networks, and finally whole "societies."

Individuals

In disciplines that focus on understanding individual behavior and the associated heterogeneity, such as cognitive psychology or behavioral economics, high-throughput virtual lab experiments can resolve scientific questions that cannot be answered by small-scale laboratory experiments designed to discriminate between specific theories. In particular, high-throughput online experiments can explore a wider range of conditions, manipulate more factors simultaneously, estimate the form of relationships with higher resolution, and support the use of more naturalistic stimuli. We will consider each of these properties in turn.

Typical behavioral experiments run with a small number of conditions or a constrained set of stimuli. For example, a researcher studying human decision-making might carefully select pairs of gambles that discriminate between existing theories and focus data collection on those gambles (e.g., (Kahneman and Tversky 1979)). Selectivity is sensible when only small numbers of participants are available — those limited resources have the most impact when applied to what we believe are the most informative cases. However, when many more participants are available, there is an opportunity to run many more conditions, and it is no longer necessary to focus on those that we believe a priori to be the most informative. Indeed, the best way to make discoveries that go against our default expectations is to explore the space of possible experiments in a way that is independent of our existing theories. Procedural generation of conditions and stimuli provides a way to do this. The researchers can define a set of dimensions along which the stimuli can vary, and then a random process can be used to generate the set of stimuli to be used in the experiment. For example, this approach was taken in the Choice Prediction Competitions run by Erev and colleagues (Erev et al. 2017; Plonsky et al. 2019), where human decision-making was studied by automatically generating over 100 random pairs of gambles following a predefined algorithm. Since each gamble can be described by a small number of dimensions that determine the probabilities and payoffs involved, this amounts to randomly sampling in the space of stimuli. Bourgin et al. (2019) took advantage of the larger sample sizes that can be obtained through virtual labs to scale up this approach, collecting human decisions for over 10,000 pairs of gambles. The resulting data set can be used to evaluate models of decision-making and is at a scale where machine learning methods can be used to augment the insights of human researchers (Agrawal, Peterson, and Griffiths 2020).

In some domains, it is possible to identify a set of factors that are all believed to influence the behavior but where the relative contributions of those factors and the ways in which they interact remain an open question. This can arise as a consequence of the hypothesis-driven approach underlying traditional behavioral research, in which different research groups each have their own hypothesis about a factor that they believe is relevant to the behavior and construct an experiment to provide evidence for that hypothesis, but the experiments provide no way to compare the factors against one another. A massively multi-factor experiment design collects all of these factors together and then runs a single experiment in which they are pitted against one another. This approach can also be extended to include interactions between the factors. The advantage, of course, by running this as part of one large, single experiment, is the ability to control for the many unmeasurable differences across studies that are likely to exist.

In other settings, the range of conditions or stimuli is intrinsically constrained, expressible in terms of just a couple of variables, but the fundamental question is about the form of the relationship between those variables. For example, the controversy over the existence of the Dunning-Kruger effect (in which people who know less about a topic significantly overestimate their knowledge) focuses on one portion of the curve that relates people's performance to their self-assessment (Kruger and Dunning 1999). If low-performers really are less metacognitively aware, as opposed to simply adjusting their estimates of their own ability in the face of uncertainty in an appropriately Bayesian way (Moore and Healy 2008), then we

should expect to see an asymmetry in the function that relates performance to self-assessment. However, previous research has not provided sufficient resolution to capture the form of this curve — the conclusion that the Dunning-Kruger effect exists was based on the aggregated performance of the lowest quartile of participants. With much larger samples, it is possible to definitively identify the form of these functions in a way that differentiates between current theories and provides the basis for future theorizing. In the case of the Dunning-Kruger effect, Jansen and colleagues used an online study with approximately 4000 participants to estimate the form of the function across the full range of performance, showing that it displays the predicted asymmetry (Jansen, Rafferty, and Griffiths 2021).

Another tradition of laboratory experiments is to use highly simplified stimuli (Rust and Movshon 2005). Such stimuli are useful for testing theories of human behavior precisely because they eliminate possible confounds. However, simplification also risks decoupling laboratory results from the naturalistic behaviors that motivated the research, sacrificing ecological validity. With much larger samples, it is possible to conduct experiments that produce meaningful results with more naturalistic stimuli. For example, the psychological literature on categorization has typically used stimuli that vary on only a handful of dimensions, such as sinusoidal gratings that vary in angle and spatial frequency or simple geometric shapes. This makes sense in data-limited settings, where categorization judgments can be collected for only a small number of stimuli so those stimuli need to vary along only a few dimensions. However, with the capacity to collect more data comes the potential to work with more complex stimuli. One recent paper evaluated computational models of categorization on a data set consisting of 10,000 natural images (Battleday et al., 2020). In order to do so, it was necessary to run an online experiment that collected approximately 500,000 human categorization decisions for those images.

Finally, as noted earlier, experiments in psychology, cognitive science, and behavioral economics traditionally study behaviors that manifest themselves on short timescales. For example, one experiment might study how the potential to establish a reputation increases participants' propensity to cooperate in a sequence of social dilemmas, while another might study how participants' critical reasoning skills correlate with their susceptibility to misinformation after reading a handful of real and fake news headlines, and another might measure their stated intention to save for their retirement after watching an image of their face being aged artificially. In reality, however, every human comprises a large collection of traits — intelligence and prosociality and forward thinkingness and social perceptiveness and many others — and exhibits behaviors that evolve and generate outcomes over their life-course. Ideally, therefore, individual-level experiments should have two properties. First, the same individuals would participate in many experiments, potentially run by different researchers at different times, and information from all these experiments should be linked so that correlations in traits and behaviors can be studied across contexts (Peysakhovich, Nowak, and Rand 2014). Second, by running individual experiments over long intervals of time (e.g., months or years) or by running repeated experiments, the relationship between traits, behavior, and outcomes can be studied on meaningful timescales. These factors both improve the validity and generalizability of experimental results as they help make experimental settings more representative of those outside the lab.

Groups

Moving up the unit of analysis from individuals to groups, new questions emerge that are not answerable even with a definitive understanding of individual behavior. For example, questions related to whether — and under what conditions — groups outperform the best individual as well as what determines collective performance have motivated innumerable studies in management and organizational science, social psychology, sociology, complexity science, and computer science. As with findings about individuals, however, research on groups has generated inconsistent and even contradictory findings. For instance, while some studies find that groups dramatically outperform individuals, others find that "process losses" cause groups to underperform their best members. Moreover, while some studies have emphasized the importance of individual skill in determining group performance, others have emphasized factors such as social perceptiveness and diversity. Faced with these conflicting findings, a hypothetical

manager would have difficulty deciding on when to assign a team to a task — versus, say, her best individual worker — how to combine individuals with different attributes, and how her decisions depend on the type and complexity of the task at hand (Richard Hackman 2011).

To be clear, the problem is not that we lack theoretically informed hypotheses about the causes and predictors of team performance. Quite to the contrary, the aforementioned literatures contain dozens to hundreds of such hypotheses, along with hundreds to thousands of empirical and modeling results. The problem is that it is unclear (a) how all these potentially relevant effects jointly predict performance; and (b) how their relative importance and interactions change over the "feature space" of tasks. In other words, our collective theoretical knowledge of what drives group performance suffers from essentially the same problem as our knowledge of individual-level decision making: isolated tests of individual theories conducted under different conditions, with different participant pools, and in different parts of the feature space, tell us little about how our theories fit together. Therefore, they have little to say about practical applications in which many theoretically motivated effects may be relevant simultaneously.

Similar remarks could be made for related classes of small-group processes, such as prosocial behavior (Fehr and Fischbacher 2003), political deliberation (Mutz 2006), and collective intelligence (Woolley et al. 2010). In all cases, there are large theoretical literatures containing many interesting hypotheses about how specific features or effects might lead to more or less cooperative behavior, more or less consensus and/or mutual understanding, and more or less intelligent decisions, among a group of people. In all cases, the corresponding empirical literature, while extensive, is filled with inconsistent or contradictory findings. And in both cases, the sum of all relevant knowledge is not well suited to directly answer straightforward questions of practical interest: "How should I go about increasing cooperative behavior in my community?" How do I change the culture of my organization to be more prosocial?" "How can I leverage the power of deliberation to increase the perceived legitimacy of local government, or to reduce polarization in society?"

These are all questions about which of many possible "levers" some policymaker or change agent might "pull" with the intent of improving some outcome of interest. Critically, however, (a) there are many such levers, (b) their individual effects may vary in magnitude and direction with other features of the problem domain and group identity, and (c) when more than one lever is pulled simultaneously or in sequence, their effects may interact with each other in important ways.

A legacy of the traditional lab model is that researchers typically identify one or a few theoretical factors of interest, and focus their experiment on the influence of those factors on some outcome behavior. Selectivity in conditions to be considered is sensible when only small numbers of participants are available. However, when many more participants are available, there is an opportunity to run many more conditions, and it is no longer necessary to focus on those that researchers believe a priori to be the most informative. In principle, researchers can define a set of dimensions along which the experiment can vary, and then a process can be used to generate and sample the set of conditions to be used in the experiment (Letham et al. 2019; Balietti, Klein, and Riedl 2018; McClelland 1997).

Therefore, answering questions of this sort requires a similar approach to that outlined above for individual behavior — i.e., some combination of procedural generation, massively multi-factor designs, high resolution, and naturalistic stimuli — except that in this case, a treatment unit is now a group rather than an individual.

Networks

The question of how influence, and more generally information, propagates in networks is pervasive throughout the social sciences and more recently in computer science. The reason is that directly influencing people to change their opinions or behavior is hard; thus, if it were possible to directly influence just a small number of people and then harness naturally occurring processes of social influence to influence some much larger number of people indirectly, such an ability would be of immense practical and policy importance (see, e.g., (Watts 2003; Christakis and Fowler 2009; Frank 2020)).

Overwhelmingly, research on social influence propagation has been theoretical in nature: some model of contagion is proposed and its behavior is then studied analytically or with the use of simulations on one or more networks, which may also be generated by a model or may come from empirical data. This approach has been extremely fruitful from a theoretical standpoint, generating numerous insights with respect to conditions under which large-scale propagation can be expected to occur (Watts 2002); differences in expected propagation for different models of contagion (D. Centola and Macy 2007); and strategies for "seeding" contagion processes so as to maximize expected propagation (Kempe, Kleinberg, and Tardos 2003). Unfortunately, the practical difficulty of running networked contagion experiments has meant that empirical tests of these theoretical models are rare (Kearns et al. 2009; D. Centola 2010; Damon Centola 2011; Mason and Watts 2012; Shore, Bernstein, and Lazer 2015; A. J. Stewart et al. 2019). Moreover, these experiments involve no more than a few tens of participants; thus, they are insufficiently large scale to generalize to most applications of interest. As a result, we have many theoretical models of contagion on networks and few experimental results that can guide real-world interventions designed to harness (or alternatively, to suppress) social influence and contagion. Having the ability to run controlled networked contagion experiments at the scale of thousands or tens of thousands of participants would revolutionize our understanding of influence maximization and prediction across many different network structures.

Societies

At its most ambitious, large-scale and high-throughput experimentation offers a new opportunity to social science: running experiments at the scale of societies. Previously, researchers who wanted to run experiments involving the interaction of hundreds of thousands of people only had the opportunity to do so in the context of field experiments. While this approach to experimentation is valuable for providing a naturalistic setting, it has a major weakness in that such experiments are hard or impossible to replicate. Furthermore, statistical analysis of complex social phenomena can be challenging because a standard field experiment provides only a single sample.

To address these limitations, social scientists have sometimes used a different tool: agent-based modeling, in which simulated agents are used as proxies for human behavior and the consequences of specific manipulations are evaluated in silico. The recruitment tools and software modules that we envision offer a third option: running behavioral simulations of the kind that are associated with agent-based models, but replacing those simulated agents with human participants (Kazerooni, Wherry, and Bazarova 2018). By constructing virtual labs that allow experimenters to design complex control structures in which one person's decision influences the decisions of others , it is possible to create simulated social systems at society scale. Crucially, the dynamics of these social systems can be replicated simply by recruiting another group of participants, opening the door to experimentally addressing a set of questions that were previously beyond the scope of social science, which can dramatically improve the validity and predictive power of results from experiments of this type. Preliminary examples of experiments that do exactly this, going so far as to simulate evolutionary processes in human populations, are already beginning to be conducted (Morgan, Suchow, and Griffiths 2020b, [a] 2020).

Challenges to achieving the promise

If the experiments just discussed are already technically possible, why are we not already running them? Why is it that, with a handful of exceptions, the vast majority of lab experiments conducted in online environments maintain the designs we are familiar with from physical labs: simple treatments, applied to just one or a few individuals, over short timescales? Why is that, with rare exceptions, the potential for virtual labs to support "high-throughput" style experiments — in which an entire parameter space can be explored with hundreds or thousands of individual experiments — remains unrealized? Why is it, in other words, that in spite of the revolutionary potential of virtual labs — a potential that has been evident for well over a decade — the main effect of virtual labs has been simply to produce somewhat more of the same, with all the same problems for replicability, robustness, and external validity? In this section, we propose four possible explanations: logistical and administrative challenges to recruiting participants; inadequately flexible and powerful software infrastructure; insufficient coordination among researchers; and cultural inertia among the research community.

Participant recruiting

Although crowdsourcing services have reduced the cost and difficulty of recruiting participants, the most popular platforms such as Amazon Mechanical Turk were designed for simple labeling tasks that can be completed independently and with little effort by individual workers about whom little is known and who vary widely in quality and persistence on the service. As a result, they are in some respects poorly suited for behavioral experiments where researchers may care about the identity of the participants, require them to devote considerable time and effort to a single task, or to interact and/or coordinate with other participants on the same task. For example:

- Although crowd workers tend to be more diverse than college students, they are not representative of the population at large (Berinsky, Huber, and Lenz 2012), and international coverage is inconsistent (Difallah, Filatova, and Ipeirotis 2018). Thus, cross-national comparative studies or studies of specific populations remain challenging to run.
- Although the major crowdsourcing services screen workers for quality, the main purpose is to filter out fraudulent and automated accounts. Thus, it continues to be difficult to identify workers with specific abilities or attributes that may be required for specific studies.
- Although crowdsourcing services provide some limited information about workers, linking individual-level data across experiments is not supported, nor is it clearly permitted by existing terms of use agreements.
- Although crowdsourcing platforms advertise large populations of workers, the pool of active workers available on any given day is typically much smaller than the total capacity, especially when restricting to high-quality workers (N. Stewart et al. 2015). Running large-scale experiments involving, say, tens of thousands of participants, remains impossible.
- Although researchers have found creative ways to arrange for multiple workers to be present simultaneously, platforms do not natively support simultaneous recruitment; thus running synchronous experiments with more than a few dozen participants remains extremely challenging.

In recent years, services such as Prolific have been introduced that adapt the crowd work model to accommodate the special needs of behavioral research; for example, Prolific offers researchers more control over participant sampling and quality as well as recruiting participants who are intrinsically motivated to contribute to scientific studies. However, no existing crowd platform yet supports the scale and throughput of human subjects required for the designs described in the previous section.

One-off infrastructure

The current state of experimental software relies on separate packages, tools, and frameworks across a wide variety of disciplines to meet specific design needs. This is inefficient because it is not only the result of much-repeated work to achieve (at least in some sense) similar goals, but because it also inhibits the ability to replicate and extend existing experiments due to incompatible software and

methodology.

While early online experiments often required extensive up-front customized software development, a number of virtual lab software packages and frameworks (e.g., Volunteer Science, nodeGame, Breadboard, Pushkin, Dallinger, jsPsych, O-Tree, LIONESS, Empirica) have now been developed that reduce the overhead associated with building and running experiments (Horton, Rand, and Zeckhauser 2011). As a result, it is now much easier to implement designs in which dozens of individuals interact synchronously in groups (Mao et al. 2016) or via networks (Shore, Bernstein, and Lazer 2015; A. J. Stewart et al. 2019; A. Almaatouq, Noriega-Campero, et al. 2020), potentially comprising a mixture of human and algorithmic agents (Shirado and Christakis 2017). It is also increasingly straightforward to implement complex research designs involving, for example, block-randomization of participants to many treatment conditions (A. Almaatouq, Yin, and J. 2020).

Existing systems, however, all exhibit trade-offs between generality and ease of deployment. This is the consequence of aiming to develop tightly integrated "end-to-end" solutions for some particular class of problems (e.g., psychology experiments, two-player economic games).

Coordination & collaboration

A third barrier to massively increase the scale, speed, and complexity of experimentation is the administrative and logistical burden of implementing these designs.

Administrative Overhead. Supervising experiments, managing subject payment and welfare, performing real-time data analysis to optimize learning, and adjusting designs in response are all labor-intensive tasks requiring varying levels of expertise. In addition, running large-scale collaborative studies requires coordination of ethics review across institutions, which leads to different data management protocols and privacy laws, in a way that poses a logistical challenge.

Mass Collaboration. As with contemporary replication studies (Open Science Collaboration 2015), major advances will probably require coordination among many labs, albeit with a different emphasis than replication. But while the "many labs" collaborations provide encouraging examples (Open Science Collaboration 2015; Salganik et al. 2020; Landy et al. 2020) , they remain relatively rare in the social sciences, which have historically rewarded individual contributions rather than team efforts. Additionally, researchers may not want to reveal their studies to the broader internet, for fear of being scooped, which would hinder sharing research designs and questions across labs (Law et al. 2017).

Cultural inertia

Twentieth-century experimental behavioral and social science evolved under a particular set of physical and logistical constraints that restricted experiments to small sample sizes, short timescales, and simple designs (e.g., testing a single hypothesis at a time). Over time, generations of researchers have internalized these features to such an extent that they are thought to be inseparable from sound scientific practice. Even as the original physical and logistical constraints are being relaxed, enabling radically different designs than in the past, these socially and culturally reified beliefs continue to shape researchers' imaginations and incentives. Thus, even as researchers in psychology and economics are recruiting ever-larger numbers of participants online, the studies that are run online differ quantitatively but not qualitatively from those that might be run in the laboratory — they use the same kind of experimental designs, with relatively small numbers of conditions, aiming to answer discrete questions that might help to differentiate between specific theories. As the scale of online experiments increases, a different approach is required: we don't want to run the same experiment with 100,000 people that we might have run with 100.

Roadmap to unlock the potential of virtual labs

Although the problems just outlined are distinct, they are also interdependent. Simply improving the usability or flexibility of experiment management software may have little marginal impact on scientific progress without the ability to run experiments in a "high throughput" manner; and that, in turn, may be impossible without much larger panels than are currently available and resolving the coordination difficulties inherent in mass collaborations. Solving all of these problems simultaneously therefore requires a level of coordinated planning and investment in shared resources that is unusual in the social sciences. In this section, we sketch out a plan to dramatically improve the scale, speed, and robustness of experimental SBE science.

New research infrastructure

In contrast with the experimental software described earlier, which has attempted to develop tightly integrated "end-to-end" solutions for some particular class of problems (e.g., psychology experiments, two-player games), an ecosystem approach involves an ensemble of functional components that are modular, interoperable, and reusable. Achieving this would require developing a set of open standards that defines what this encapsulation (service/component) is, how to communicate with it, and how to find and use it.

While many of the existing software platforms have instantiated some of these functional components (although in tightly integrated applications designed to meet specific experimental needs), some components will need to be substantially expanded in scope and functionality in order to scale experimental SBE science.

The use of the "ecosystem" as a design principle presents several opportunities for operational efficiency.

1. An ecosystem will allow, at least in theory[1], for the reuse of current software assets, in turn lowering new development costs, decreasing development time, reducing risk, and leveraging existing platform investments and strengths.

2. The individual components of the ecosystem should be loosely coupled to reduce vendor/provider lock-in and create a flexible infrastructure. As a result, the individual components of the ecosystem should be modular in the sense that each can be modified or replaced without needing to modify or replace any other component because the interface to the component remains the same. The resulting functional components will be available for end-users (i.e., researchers) to amalgamate (or mashup) into situational, creative, and novel experiments in ways that the developers may not originally envision. Indeed it is precisely because no one particular platform, as they now exist, can be expected to offer optimal functionality for all experimental designs, that we believe a modular design is necessary.

3. The functional scope of these components should allow for the possibility to directly define experiment requirements as a collection of these functional components, rather than translating experiment requirements into lower-level software development requirements. As a result, the ecosystem should abstract away many of the logistical concerns of running experiments, analogous to how cloud computing has abstracted away from the management of technical resources for many companies.

[1]Unfortunately, most existing frameworks were not, in general, designed as a collection of well-encapsulated components to be externally shared and used by other applications, let alone to meet higher-level and generally more abstract experimental design requirements.

New Experimental Designs

The ability to conduct procedurally generated, massively multi-factor, high-resolution, naturalistic designs will change the way that we approach running behavioral experiments. However, there is still a lot of room to develop other kinds of experimental designs that are optimized for the high-throughput environment created by virtual labs. In particular, we can navigate the increasingly large spaces of possible conditions and stimuli supported by online experiments by making use of adaptive designs that intelligently determine the next conditions to run.

Adaptive designs leverage large samples by conducting an experiment that is dynamic rather than static. The traditional behavioral research paradigm, in which participants would spend a significant portion of an hour answering questions in a single experimental condition, is an inefficient way of answering questions about human behavior. Using experiments that are administered by computer, in which the responses of one participant can influence the questions that are asked of

another, it is possible to draw upon methods from computer science and statistics to construct designs that can create a more complete picture of aspects of human cognition (Suchow and Griffiths 2016).

Some of the most ambitious adaptive designs essentially implement an algorithm with people. For example, iterated learning designs, in which each participant learns from data generated by the previous participant (Kalish and Griffiths 2006), can be viewed as a kind of Markov chain Monte Carlo algorithm implemented with people. This kind of design can be an effective way of revealing the structure of human learning biases. Algorithmic designs may be more challenging to develop collaboratively, but also offer disproportionate benefits in allowing us to make the most of the large sample sizes provided by our experiments.

More generally, automating the administration of experiments creates opportunities for automating experiment design. If experiments are run by specifying a set of parameters and then executing a line of computer code, we can write algorithms that automatically decide how to set those parameters. For example, if the goal is to identify the conditions that give rise to the biggest change in behavior, we can express that as an objective function and run an optimization algorithm over the parameters of an experiment, executing variance of the experiment to determine how people behave. Alternatively, if the goal is to estimate a theoretical model (or choose between models) we can automate the process of identifying the experiment parameters that are most informative, that iteratively run those experiments to hone in on the answer to the question at hand.

New approaches to data analysis

The data sets produced by large-scale studies can pose a challenge for the traditional methods of analysis and modeling used in behavioral research. Many disciplines that use experiments rely on statistical significance testing as a means of evaluating hypotheses. However, as data set sizes increase statistical significance becomes less meaningful — at the significance levels traditionally used in most social science research, even the smallest effect will be significant. This encourages developing a new set of conventions for the analysis of large-scale data sets, supporting the use of more stringent alpha levels, an emphasis on effect sizes, multilevel models, and Bayesian methods — all practices that have been encouraged in recent papers on improving methodology in social science research.

More fundamentally, the style of procedurally generated, massively multi-factor, high-resolution, naturalistic designs that we have advocated for corresponds to a very different approach to theory testing than traditional null hypothesis testing. Returning to our analogy of pulling levers (see 2.3), null hypothesis testing poses and answers questions of the form “Can I reject the null hypothesis that lever X has no effect on the outcome?” Setting aside that some of these results may fail to replicate, simply showing that many “levers” have non-zero effects under some conditions tells us little about the practical question of which levers to pull and how much we can expect them to affect the outcome. But shifting to an approach in which many levers are tested simultaneously and all these multi-factor experiments are then replicated over many different conditions, potentially in an adaptive manner, will require embracing (or at least accepting) a very different notion of what is considered a theoretical and empirical contribution (e.g. a null effect may be of significant interest when done in large scale).

Large-scale data sets also potentially pose a challenge for computational modeling, in part because as the size of the data set increases so does the complexity of the model that it supports. A principle of statistics known as the bias-variance trade-off identifies two ways in which a model can fail to generalize: it can be too simple, in which case it is unable to capture the trends in the data and exhibits a bias, or it can be too complex, in which case it overfits the data and shows a lot of variance across data sets (Geman, Bienenstock, and Doursat 1992). However, this variance decreases as the data sets increase in size. Consequently, with large data sets poor generalization is typically due to models being too simple rather than being too complex.

The implications of the bias-variance trade-off for social science are perhaps surprising: as we run larger experiments, the simple models that we use to explain behavior are no longer going to be adequate. The data are going to show systematic regularities that go beyond the capacity of any previous models to explain. As a consequence, we need to develop new kinds of models that are

capable of dealing with the complexity of human behavior while retaining the interpretability of simpler models. An early example of this approach is Agrawal et al. (Agrawal, Peterson, and Griffiths 2020), who combine machine learning models with rational choice models to jointly maximize predictive accuracy and interpretability in the context of moral judgments.

New kinds of theories

The implications of the bias-variance trade-off run deep. If individuals, groups, and societies are complex, then as we get more data we are going to reveal more and more of that complexity. This creates a challenge for disciplines that have tried to reduce human behavior to simple theories. In many ways, the fact that we have been able to do so in the past is partly a consequence of the fact that we didn't have enough data. In parallel with the changes to infrastructure and methodology required to fulfill the promise of virtual labs, we as scientists need to adjust our expectations about what theories of human behavior are going to look like. We will discover powerful new general laws that characterize what people do across many different situations — something that is only possible when you are able to study all of those situations — but we will also discover all of the complexities and nuances of the ways that people deviate from those general laws, and have enough data to establish that those deviations are systematic and meaningful. Our new theories are going to let us predict human behavior better than ever before, and understand the factors that influence it, but at the cost of some of the simplicity that our limited data have led us to believe in

Conclusion

Returning to our opening motivation, while experimental social, behavioral, and economic science has clearly benefited from the digital revolution, it has developed more slowly and accomplished less than it could have. In this paper we have reviewed some of the exciting progress made to date but also argued that truly transformative progress will require more than just business as usual. Specifically, we have outlined a series of needed developments that we believe will, in combination, substantially improve the robustness, replicability, and ultimately usefulness of lab-style experiments applied to problems of human behavior, economics, and society. Although ambitious, the technical and financial investments required to execute this plan are modest in comparison to the potential scientific impact it can have. By developing a new mentality of virtual lab infrastructure a flood of new designs and novel experiments can become available, helping to resolve many of the challenges experimental social scientists have faced so far.

References

Agrawal, Mayank, Joshua C. Peterson, and Thomas L. Griffiths. 2020. “Scaling up Psychology via Scientific Regret Minimization.” Proceedings of the National Academy of Sciences of the United States of America 117 (16): 8825–35.

Almaatouq, Abdullah, Joshua Becker, James P. Houghton, Nicolas Paton, Duncan J. Watts, and Mark E. Whiting. 2020. “Empirica: A Virtual Lab for High-Throughput Macro-Level Experiments.” arXiv [cs.HC]. arXiv. http://arxiv.org/abs/2006.11398.

Almaatouq, Abdullah Mohammed. 2019. “Towards Stable Principles of Collective Intelligence under an Environment-Dependent Framework.” Massachusetts Institute of Technology. https://dspace.mit.edu/handle/1721.1/123223?show=full.

Almaatouq, Abdullah, Alejandro Noriega-Campero, Abdulrahman Alotaibi, P. M. Krafft, Mehdi Moussaid, and Alex Pentland. 2020. “Adaptive Social Networks Promote the Wisdom of Crowds.” Proceedings of the National Academy of Sciences of the United States of America 117 (21): 11379–86.

Almaatouq, Abdullah, Ming Yin, and Watts Duncan J. 2020. “Collective Problem-Solving of Groups Across Tasks of Varying Complexity.” 10.31234/osf.io/ra9qy. PsyArXiv. https://psyarxiv.com/ra9qy.

Asch, Solomon E. 1951. “Effects of Group Pressure upon the Modification and Distortion of Judgments.” Organizational Influence Processes, 295–303.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. “The Moral Machine Experiment.” Nature 563 (7729): 59–64.

Balietti, Stefano, Brennan Klein, and Christoph Riedl. 2018. “Fast Model-Selection through Adapting Design of Experiments Maximizing Information Gain.” arXiv [stat.AP]. arXiv. http://arxiv.org/abs/1807.07024.

Battleday, R.M., Peterson, J.C. and Griffiths, T.L., 2020. Capturing human categorization of natural images by combining deep networks and cognitive models. Nature communications, 11(1), pp.1-14.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. “Evaluating Online Labor Markets for Experimental Research: Amazon.com’s Mechanical Turk.” Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association 20 (3): 351–68.

Bourgin, David D., Joshua C. Peterson, Daniel Reichman, Stuart J. Russell, and Thomas L. Griffiths. 2019. “Cognitive Model Priors for Predicting Human Decisions.” In Proceedings of the 36th International Conference on Machine Learning, edited by Kamalika Chaudhuri and Ruslan Salakhutdinov, 97:5133–41. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR.

Centola, D. 2010. “The Spread of Behavior in an Online Social Network Experiment.” Science 329 (5996): 1194.

Centola, Damon. 2011. “An Experimental Study of Homophily in the Adoption of Health Behavior.” Science 334 (6060): 1269–72.

Centola, D., and M. Macy. 2007. “Complex Contagions and the Weakness of Long Ties.” AJS; American Journal of Sociology 113 (3): 702–34.

Christakis, Nicholas A., and James H. Fowler. 2009. Connected: The Surprising Power of Social Networks and How They Shape Our Lives. New York: Little Brown.

Difallah, D., E. Filatova, and P. Ipeirotis. 2018. “Demographics and Dynamics of Mechanical Turk Workers.” Conference on Web Search and Data .... https://dl.acm.org/doi/abs/10.1145/3159652.3159661?casa_token=pK-1rI4ryT8AAAAA:wMGlzEHkVLsz7BrAUhEZb-qlEa0g3JUCJSaCV51bLBcfKLhwXVdb9-cyksZrCOSzkLmWyWduJEnSZA.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. “Chapter 61 Using Randomization in Development Economics Research: A Toolkit.” In Handbook of Development Economics, edited by T. Paul Schultz and John A. Strauss, 4:3895–3962. Elsevier.

Embrey, M., G. R. Fréchette, and S. Yuksel. 2018. “Cooperation in the Finitely Repeated Prisoner’s Dilemma.” The Quarterly Journal of. https://academic.oup.com/qje/article-abstract/133/1/509/4095199.

Erev, Ido, Eyal Ert, Ori Plonsky, Doron Cohen, and Oded Cohen. 2017. “From Anomalies to Forecasts: Toward a Descriptive Model of Decisions under Risk, under Ambiguity, and from Experience.” Psychological Review 124 (4): 369–409.

Fehr, Ernst, and U. Fischbacher. 2003. “The Nature of Human Altruism.” Nature 425 (October): 785–91.

Frank, Robert H. 2020. “Under the Influence.” https://doi.org/10.2307/j.ctvmd85w3.

Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. “Neural Networks and the Bias/Variance Dilemma.” Neural Computation 4 (1): 1–58.

Goroff, Daniel L., Jr Lewis Neil A, Anne M. Scheel, Laura D. Scherer, and Joshua A. Tucker. 2018. “The Inference Engine: A Grand Challenge to Address the Context Sensitivity Problem in Social Science Research.” https://doi.org/10.31234/osf.io/j8b9a.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. “In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies.” The American Economic Review 91 (2): 73–78.

———. 2005. “‘Economic Man’ in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies.” The Behavioral and Brain Sciences 28 (6): 795–815.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. “The Weirdest People in the World?” The Behavioral and Brain Sciences 33 (2-3): 61–83.

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. “The Online Laboratory: Conducting Experiments in a Real Labor Market.” Experimental Economics 14 (3): 399–425.

Jansen, Rachel A., Anna N. Rafferty, and Thomas L. Griffiths. 2021. “A Rational Model of the Dunning-Kruger Effect Supports Insensitivity to Evidence in Low Performers.” Nature Human Behaviour, February. https://doi.org/10.1038/s41562-021-01057-0.

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk.” Econometrica. https://doi.org/10.2307/1914185.

Kalish, Michael L., and Thomas L. Griffiths. 2006. “Iterated Learning Reveals Inductive Priors in Function and Category Learning.” PsycEXTRA Dataset. https://doi.org/10.1037/e527352012-177.

Kazerooni, F., O. D. Wherry, and N. N. Bazarova. 2018. “Upstanding by Design: Bystander Intervention in Cyberbullying.” CHI ... Conference Proceedings / Conference on Human Factors in Computing Systems. CHI Conference. https://dl.acm.org/doi/abs/10.1145/3173574.3173785?casa_token=l54ac6w0aF4AAAAA:St9XBUuMzILAeh7uajLsgX8218UL4_j-yLonGz-GzlKWp6HzppXHa-vPL6zOpMGHEEp-CmIaSSSiBQ.

Kearns, M., S. Judd, J. Tan, and J. Wortman. 2009. “Behavioral Experiments on Biased Voting in Networks.” Proceedings of the National Academy of Sciences 106 (5): 1347.

Kempe, David, Jon Kleinberg, and Eva Tardos. 2003. Maximizing the Spread of Influence through a Social Network. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA.: Association of Computing Machinery.

Kirby, Simon, Hannah Cornish, and Kenny Smith. 2008. “Cumulative Cultural Evolution in the Laboratory: An Experimental Approach to the Origins of Structure in Human Language.” Proceedings of the National Academy of Sciences of the United States of America 105 (31): 10681–86.

Kohavi, Ron, Roger Longbotham, and Toby Walker. 2010. “Online Experiments: Practical Lessons.” Computer, 2010.

Kruger, J., and D. Dunning. 1999. “Unskilled and Unaware of It: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self-Assessments.” Journal of Personality and Social Psychology 77 (6): 1121–34.

Landy, Justin F., Miaolei (liam) Jia, Isabel L. Ding, Domenico Viganola, Warren Tierney, Anna Dreber, Magnus Johannesson, et al. 2020. “Crowdsourcing Hypothesis Tests: Making Transparent How Design Choices Shape Research Results.”

Psychological Bulletin. https://doi.org/10.1037/bul0000220.

Law, Edith, Krzysztof Z. Gajos, Andrea Wiggins, Mary L. Gray, and Alex Williams. 2017. “Crowdsourcing as a Tool for Research: Implications of Uncertainty.” In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 1544–61. CSCW ’17. New York, NY, USA: Association for Computing Machinery.

Letham, Benjamin, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. 2019. “Constrained Bayesian Optimization with Noisy Experiments.” Bayesian Analysis. https://doi.org/10.1214/18-ba1110.

Manzi, Jim. 2012. “Uncontrolled: The Surprising Payoff of Trial-and-Error for Business.” Politics, and Society. Basic Books, 1–320.

Mao, Andrew, Lili Dworkin, Siddharth Suri, and Duncan J. Watts. 2017. “Resilient Cooperators Stabilize Long-Run Cooperation in the Finitely Repeated Prisoner’s Dilemma.” Nature Communications 8: 13800.

Mao, Andrew, Winter Mason, Siddharth Suri, and Duncan J. Watts. 2016. “An Experimental Study of Team Size and Performance on a Complex Task.” PloS One 11 (4): e0153048.

Mason, W., and S. Suri. 2012. “Conducting Behavioral Research on Amazon’s Mechanical Turk.” Behavior Research Methods 44 (1): 1–23.

Mason, W., and D. J. Watts. 2012. “Collaborative Learning in Networks.” Proceedings of the National Academy of Sciences 109 (3): 764–69.

McClelland, Gary H. 1997. “Optimal Design in Psychological Research.” Psychological Methods 2 (1): 3–19.

Milgram, S. 1969. Obedience to Authority. New York: Harper and Row.

Moore, Don A., and Paul J. Healy. 2008. “The Trouble with Overconfidence.” Psychological Review 115 (2): 502–17.

Morgan, Thomas J. H., Jordan W. Suchow, and Thomas L. Griffiths. 2020a. “What the Baldwin Effect Affects Depends on the Nature of Plasticity.” Cognition 197 (April): 104165.

———. 2020b. “Experimental Evolutionary Simulations of Learning, Memory and Life History.” Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 375 (1803): 20190504.

Musch, Jochen, and Ulf-Dietrich Reips. 2000. “A Brief History of Web Experimenting.” Psychological Experiments on the Internet. https://doi.org/10.1016/b978-012099980-4/50004-6.

Mutz, Diana C. 2006. Hearing the Other Side: Deliberative versus Participatory Democracy. Cambridge University Press.

Nisbett, Richard. 2004. The Geography of Thought: How Asians and Westerners Think Differently...and Why. Simon and Schuster.

Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” Science 349 (6251): 10.1126/science.aac4716.

Paolacci, Gabriele, Jess Chandler, and Panos G. Ipeirotis. 2010. “Running Experiments on Amazon Mechanical Turk.” Judgment and Decision Making 5 (5): 411–19.

Peysakhovich, Alexander, Martin A. Nowak, and David G. Rand. 2014. “Humans Display a ‘cooperative Phenotype’ That Is Domain General and Temporally Stable.” Nature Communications. https://doi.org/10.1038/ncomms5939.

Plonsky, Ori, Reut Apel, Eyal Ert, Moshe Tennenholtz, David Bourgin, Joshua C. Peterson, Daniel Reichman, et al. 2019. “Predicting Human Decisions with Behavioral Theories and Machine Learning.” arXiv [cs.AI]. arXiv. http://arxiv.org/abs/1904.06866.

Plott, Charles R. 1982. “Industrial Organization Theory and Experimental Economics.” Journal of Economic Literature 20 (4): 1485–1527.

Reips, Ulf-Dietrich. 2012. “Using the Internet to Collect Data.” https://psycnet.apa.org/record/2011-23864-017.

Richard Hackman, J. 2011. Collaborative Intelligence: Using Teams to Solve Hard Problems. Berrett-Koehler Publishers.

Rust, Nicole C., and J. Anthony Movshon. 2005. “In Praise of Artifice.” Nature Neuroscience 8 (12): 1647–50.

Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. 2006. “Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market.” Science 311 (5762): 854–56.

Salganik, Matthew J., Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, et al. 2020. “Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration.” Proceedings of the National Academy of Sciences of the United States of America 117 (15): 8398–8403.

Shirado, Hirokazu, and Nicholas A. Christakis. 2017. “Locally Noisy Autonomous Agents Improve Global Human Coordination in Network Experiments.” Nature 545 (7654): 370–74.

Shore, Jesse, Ethan Bernstein, and David Lazer. 2015. “Facts and Figuring: An Experimental Investigation of Network Structure and Performance in Information and Solution Spaces.” Organization Science.

Stewart, Alexander J., Mohsen Mosleh, Marina Diakonova, Antonio A. Arechar, David G. Rand, and Joshua B. Plotkin. 2019. “Information Gerrymandering and Undemocratic Decisions.” Nature 573 (7772): 117–21.

Stewart, Neil, Christoph Ungemach, Adam J. L. Harris, Daniel M. Bartels, Ben R. Newell, Gabriele Paolacci, Jesse Chandler, and Others. 2015. “The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers.” Judgment and Decision Making 10 (5): 479–91.

Suchow, Jordan W., and Thomas L. Griffiths. 2016. “Rethinking Experiment Design as Algorithm Design.” Advances in Neural Information Processing Systems 29: 1–8.

Tversky, Amos, and Daniel Kahneman. 1974. “Judgment under Uncertainty: Heuristics and Biases.” Science 185 (4157): 1124–31.

Valentine, Melissa A., Daniela Retelny, Alexandra To, Negar Rahmati, Tulsee Doshi, and Michael S. Bernstein. 2017. “Flash Organizations: Crowdsourcing Complex Work by Structuring Crowds As Organizations.” In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 3523–37. CHI ’17. New York, NY, USA: ACM.

Watts, Duncan J. 2002. “A Simple Model of Global Cascades on Random Networks.” Proceedings of the National Academy of Sciences of the United States of America 99 (9): 5766–71.

———. 2003. Six Degrees: The Science of A Connected Age. New York: W. W. Norton.

———. 2017. “Should Social Science Be More Solution-Oriented?” Nature Human Behaviour 1: 0015.

Whiting, Mark E., Dilrukshi Gamage, Snehalkumar (neil) S. Gaikwad, Aaron Gilbee, Shirish Goyal, Alipta Ballav, Dinesh Majeti, et al. 2017. “Crowd Guilds: Worker-Led Reputation and Feedback on Crowdsourcing Platforms.” Proceedings of the. https://dl.acm.org/citation.cfm?id=2998234.

Whiting, Mark E., Irena Gao, Michelle Xing, Junior Diarrassouba N’godjigui, Tonya Nguyen, and Michael S. Bernstein. 2020. “Parallel Worlds: Repeated Initializations of the Same Team To Improve Team Viability.” CSCW. https://doi.org/10.1145/3392877.

Woolley, A. W., C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. 2010. “Evidence for a Collective Intelligence Factor in the Performance of Human Groups.” Science.

Yamagishi, Toshio. 1986. “The Provision of a Sanctioning System as a Public Good.” Journal of Personality and Social Psychology 51 (1): 110–16.

Yarkoni, Tal. 2019. “The Generalizability Crisis.” https://doi.org/10.31234/osf.io/jqw35.