Perspective

Integrating explanation and prediction in computational social science

🔴 Check for updates

Jake M. Hofman[1,17 ✉], Duncan J. Watts[2,3,4,17 ✉], Susan Athey[5], Filiz Garip[6], Thomas L. Griffiths[7,8], Jon Kleinberg[9,10], Helen Margetts[11,12], Sendhil Mullainathan[13], Matthew J. Salganik[6], Simine Vazire[14], Alessandro Vespignani[15] & Tal Yarkoni[16]

Computational social science is more than just large repositories of digital data and the computational methods needed to construct and analyse them. It also represents a convergence of different fields with different ways of thinking about and doing science. The goal of this Perspective is to provide some clarity around how these approaches differ from one another and to propose how they might be productively integrated. Towards this end we make two contributions. The first is a schema for thinking about research activities along two dimensions—the extent to which work is explanatory, focusing on identifying and estimating causal effects, and the degree of consideration given to testing predictions of outcomes—and how these two priorities can complement, rather than compete with, one another. Our second contribution is to advocate that computational social scientists devote more attention to combining prediction and explanation, which we call integrative modelling, and to outline some practical suggestions for realizing this goal.

In the past 15 years, social science has experienced the beginnings of a 'computational revolution' that is still unfolding[1–4]. In part this revolution has been driven by the technological revolution of the internet, which has effectively digitized the social, economic, political, and cultural activities of billions of people, generating vast repositories of digital data as a byproduct[5]. And in part it has been driven by an influx of methods and practices from computer science that were needed to deal with new classes of data—such as search and social media data—that have tended to be noisier, more unstructured, and less 'designed' than traditional social science data (for example, surveys and lab experiments). One obvious and important outcome of these dual processes has been the emergence of a new field, now called computational social science[2,4], that has generated considerable interest among social scientists and computer scientists alike[6].

What we argue in this paper, however, is that another outcome—less obvious but potentially even more important—has been the surfacing of a tension between the epistemic values of social and computer scientists. On the one hand, social scientists have traditionally prioritized the formulation of interpretatively satisfying explanations of individual and collective human behaviour, often invoking causal mechanisms derived from substantive theory[7]. On the other hand, computer scientists have traditionally been more concerned with developing accurate predictive models, whether or not they correspond to causal mechanisms or are even interpretable[8].

In turn, these different values have led social and computer scientists to prefer different methods from one another, and to invoke different standards of evidence. For example, whereas quantitative methods in social science are designed to identify causal relationships or to obtain unbiased estimates of theoretically interesting parameters, machine learning methods are typically designed to minimize total error on as-yet unseen data[9,10]. As a result, it is standard practice for social scientists to fit their models entirely 'in-sample', on the grounds that they are seeking to explain social processes and not to predict outcomes, whereas for computer scientists evaluation on 'held out' data is considered obligatory[11]. Conversely, computer scientists often allow model complexity to increase as long as it continues to improve predictive performance, whereas for social scientists models should be grounded in, and therefore constrained by, substantive theory[12].

We emphasize that both approaches are defensible on their own terms, and both have generated large, productive scientific literatures; however, both approaches have also been subjected to serious criticism. On the one hand, theory-driven empirical social science has been criticized for generating findings that fail to replicate[13], fail to generalize[14], fail to predict outcomes of interest[15,16], and fail to offer solutions to real-world problems[17,18]. On the other hand, complex predictive models have also been criticized for failing to generalize[19] as well as being uninterpretable[20] and biased[21]. Meanwhile, extravagant claims that the ability to mine sufficiently large datasets will result in an 'end of theory' have been widely panned[22]. How might we continue

[1]Microsoft Research, New York, NY, USA. [2]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. [3]The Annenberg School of Communication, University of Pennsylvania, Philadelphia, PA, USA. [4]Operations, Information, and Decisions Department, University of Pennsylvania, Philadelphia, PA, USA. [5]Graduate School of Business, Stanford University, Stanford, CA, USA. [6]Department of Sociology, Princeton University, Princeton, NJ, USA. [7]Department of Psychology, Princeton University, Princeton, NJ, USA. [8]Department of Computer Science, Princeton University, Princeton, NJ, USA. [9]Department of Computer Science, Cornell University, Ithaca, NY, USA. [10]Department of Information Science, Cornell University, Ithaca, NY, USA. [11]Oxford Internet Institute, University of Oxford, Oxford, UK. [12]Public Policy Programme, The Alan Turing Institute, London, UK. [13]Booth School of Business, University of Chicago, Chicago, IL, USA. [14]Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, Victoria, Australia. [15]Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University, Boston, MA, USA. [16]Department of Psychology, University of Texas at Austin, Austin, TX, USA. [17]These authors contributed equally: Jake M. Hofman, Duncan J. Watts. ✉e-mail: jmh@microsoft.com; djwatts@seas.upenn.edu

Perspective

to benefit from the decades of thinking and methodological development that have been invested in these two canonical traditions while also acknowledging the legitimacy of these criticisms? Relatedly, how might social and computer scientists constructively reconcile their distinct epistemic values to produce new methods and standards of evidence that both can agree are desirable?

Our position is that each tradition, while continuing to advance its own goals, can benefit from taking seriously the goals of the other. Specifically, we make two related contributions. First, we argue that while the goals of prediction and explanation appear distinct in the abstract they can easily be conflated in practice, leading to confusion about what any particular method can accomplish. We introduce a conceptual framework for categorizing empirical methods in terms of their relative emphasis on prediction and explanation. In addition to clarifying the distinction between predictive and explanatory modelling, this framework reveals a currently rare class of methods that integrate the two. Second, we offer a series of suggestions that we hope will lead to more of what we call integrative modelling. In addition, we advocate for clearer labelling of the explanatory and predictive power of individual contributions and argue that open science practices should be standardized between the computational and social sciences. In summary, we conclude that while exclusively explanatory or predictive approaches can and do contribute to our understanding of a phenomenon, claims to have understood that phenomenon should be evaluated in terms of both. Considering the predictive power of explanatory models can help to prioritize the causal effects we investigate and quantify how much they actually explain, and may reveal limits to our understanding of phenomena. Conversely, an eye towards explanation can focus our attention on the prediction problems that matter most and encourage us to build more robust models that generalize better under interventions and changes. Taking both explanation and prediction seriously will therefore be likely to require researchers to embrace epistemic modesty, but will advance work at the intersection of the computational and social sciences.

Prediction versus explanation

To illustrate how the goals of prediction and explanation can be conflated, consider the common practice of employing null hypothesis significance testing (NHST)[23] to reject a null hypothesis[23,24] that some theoretically motivated effect is absent (that is, is exactly zero) with a confidence that is controlled by a fixed false-positive rate, traditionally set to 5%. For example, a study might seek to reject the null hypothesis that a job applicant's perceived race has no effect on their prospects of being hired[25], or that ethnic or religious divisions within a country have no effect on the likelihood of civil war[15].

As many previous authors have noted, NHST has been widely misapplied in numerous ways—underpowered experiments, multiple comparisons, inappropriate stopping rules, and so on—that tend to produce a surprisingly high rate of false-positive findings[26,27], and have led to widely discussed replication problems[28]. From the perspective of integrating explanation and prediction, NHST is problematic for other, more fundamental reasons. NHST invokes the language of prediction; however, the prediction that is being made is often not directly about the outcome of interest, nor even about the magnitude of some theoretically interesting effect, but simply that the hypothesized effect is not zero. In other words, a common application of NHST is not so much to test predictions at all but instead to argue that a theory is not inconsistent with the data and then to use the theory as an explanatory tool. Furthermore, while there are circumstances under which it is useful to show that an effect is unlikely to be zero, in the complex world of human and social behaviour it is highly likely that many effects are non-zero[29,30]. Showing that one's preferred theory cannot be ruled out by the data is therefore an exceptionally weak test of the theory[31,32], and hence explains much less than it appears to.

Conversely, purely predictive exercises can also risk confusing prediction with explanation. Predictive models that exploit statistical associations to forecast outcomes, sometimes with seemingly impressive accuracy, can confer the feeling of having understood a phenomenon. But they often rely, sometimes implicitly, on the assumption that these predictions are to be evaluated exclusively in settings where the relationships between the predictors and the outcome of interest are stable[33]. As a result, model performance can change markedly under interventions that alter the associations in question[19], or can otherwise result in biased or misleading interpretations[34].

In fact, 'predicting an outcome' can refer to many different activities for which expectations of accuracy may vary widely. For example, the finding that the volume of influenza-related search queries in a particular geographic region is highly correlated ($r = 0.9$) with caseload data from the US Centers for Disease Control (CDC) reported two weeks later seems impressive, until it is revealed that the same correlation can be obtained directly from the CDC data alone simply by using case counts from previous weeks to forecast those for future weeks[35]. Whether a particular model is considered valuable or not therefore depends not only on its absolute performance, but also its comparison to the appropriate baseline(s).

In addition, the very same model estimated on the same data can yield qualitatively different conclusions regarding apparent predictive accuracy—ranging from 'extremely accurate' to 'relatively poor'—simply by making different choices during the evaluation procedure[36]. By analogy with NHST, not only can predictive modelling appear to generate explanations when it does not; the predictions themselves may be much weaker than they appear.

A framework for integrative modelling

As these examples illustrate, the relationship between explanation and prediction is often blurry in practice and can lead to confusion about which goals are being satisfied by any particular research activity. To clarify our thinking, we shift from talking about explanation and prediction in the abstract to more specifically discussing the types of empirical modelling activity that are common throughout computational and social science.

We emphasize that our focus here is on empirical modelling activities, not theoretical modelling such as mathematical and agent-based modelling. Theoretical work, which includes modelling as well as substantive and qualitative theory, is an essential counterpart to empirical work—for example, theory is necessary in order to identify appropriate constructs to measure or predict, or to propose hypotheses to test. Here, however, we wish to focus on research activities whose aim is to test and validate models using empirical data. To further clarify the scope of our argument, by 'models' we mostly mean the types of statistical and algorithmic models that are widely used in quantitative social science, data science, and applied machine learning. However, our framework could also be applied to explanatory and predictive analyses more generally (for example, mechanistic models, small-n case studies or comparative studies, studies using prediction markets, and so on) as long as they somehow use empirical data to validate explanations or predictions.

Concretely, we propose the conceptual framework illustrated schematically in Table 1. The two dimensions of the Table represent differing levels of emphasis placed on explanation and prediction, respectively, where we have partitioned the space into four quadrants: descriptive modelling, explanatory modelling, predictive modelling, and integrative modelling.

Descriptive modelling (quadrant 1) refers to activities that are fundamental to any scientific endeavour: how to think about, define, measure, collect, and describe relationships between quantities of interest. Activities in this quadrant include traditional statistics and survey research as well as computational methods such as topic modelling

Table 1 | A schematic for organizing empirical modelling along two dimensions, representing the different levels of emphasis placed on prediction and explanation

	No intervention or distributional changes	Under interventions or distributional changes
Focus on specific features or effects	Quadrant 1: Descriptive modelling Describe situations in the past or present (but neither causal nor predictive)	Quadrant 2: Explanatory modelling Estimate effects of changing a situation (but many effects are small)
Focus on predicting outcomes	Quadrant 3: Predictive modelling Forecast outcomes for similar situations in the future (but can break under changes)	Quadrant 4: Integrative modelling Predict outcomes and estimate effects in as yet unseen situations

The rows highlight where we focus our attention (on either specific features that might affect an outcome of interest, or directly on the outcome itself), whereas the columns specify what types of situations we are modelling (a 'fixed' world in which no changes or interventions take place, or one in which features or inputs are actively manipulated or change owing to other uncontrolled forces).

and community detection in networks[10]. For example, much of what is known about public opinion, the state of the economy, and everyday human experience is derived from survey research, whether conducted by federal statistical agencies such as the Bureau of Labour Statistics or research organizations such as Pew Research Center. Statistical analyses of administrative data are also often descriptive in nature. For example, recent studies have documented important differences in mortality rates[37], wealth gaps[38] and intergenerational economic mobility[39] across racial and ethnic groups. Qualitative and comparative methods that are popular in sociology, communications, and anthropology also fall into this quadrant. Finally, much of the progress in computational social science to date has been in using digital signals and platforms to investigate previously unmeasurable concepts[5,40]. Descriptive work, in other words, whether qualitative or quantitative, is useful and interesting in its own right and also foundational to the activities conducted in the other three quadrants.

Moving beyond description, explanatory modelling (quadrant 2) refers to activities whose goal is to identify and estimate causal effects, but that do not focus directly on predicting outcomes. Most of traditional empirical sociology, political science, economics, and psychology falls into this quadrant, which encompasses a wide range of methods, including statistical modelling of observational data, lab experiments, field experiments, and qualitative methods. Some methods (for example, in randomized or natural experiments, or non-experimental identification strategies such as instrumental variables and regression discontinuity designs) isolate causal effects by design, whereas others (for example, regression modelling, qualitative data) invoke causal interpretations based on theory. Regardless, methods in this quadrant tend to prioritize simplicity, considering one or only a handful of features that may affect an outcome of interest. We emphasize that these approaches can be very useful for understanding individual causal effects, shaping theoretical models, and even guiding policy. For example, field experiments that show that job applicants with characteristically 'Black' names are less likely to be interviewed than those with 'white' names[25] reveal the presence of structural racism and inform public debates about discrimination with respect to gender, race, and other protected attributes. Relatedly, quantifying difficult-to-assess effects, such as the impact of gender and racial diversity on policing[41], can motivate concrete policy interventions. Nonetheless, the emphasis on studying effects in isolation can lead to little, if any, attention being paid to predictive accuracy. As many effects are small, and simple models can fail to incorporate the broader set of features pertinent to the outcome being studied, these methods can suffer from relatively poor predictive performance.

In contrast with explanatory modelling, predictive modelling (quadrant 3) refers to activities that attempt to predict the outcome of interest directly but do not explicitly concern themselves with the identification of causal effects. 'Prediction' in this quadrant may or may not be about actual future events; however, in contrast with quadrants 1 and 2, it refers exclusively to 'out of sample' prediction[42], meaning that the data on which the model is evaluated (the held-out or test data) are different from the data on which the model was estimated (the training data). Activities in this quadrant encompass time series modelling[43], prediction contests[44], and much of supervised machine learning[10], ranging from simple linear regression to complex artificial neural networks. By evaluating performance on a held-out test set, these methods focus on producing predictions that generalize well to future observations. From a policy perspective, it can be helpful to have high-quality forecasts of future events even if those forecasts are not causal in nature[9,45–47]. For example, applications of machine learning to human behaviour abound in online advertising and recommendation systems, but can also detect potentially viral content on social media early in its trajectory[48]. Although these algorithms do not identify what is causing people to click or content to spread, they can still be useful inputs for decision-makers—for example, alerting human reviewers to check potentially large cascades for harmful misinformation. That said, there is often an implicit assumption that the data used to train and test the model come from the same data-generating process, akin to making forecasts in a static (albeit possibly noisy) world. As a result, while these methods often work well for a fixed data distribution, they may not generalize to settings in which features or inputs are actively manipulated (as in a controlled experiment or policy change) or change as a result of other, uncontrolled factors.

Combining the explanatory properties of quadrant 2 and the predictive properties of quadrant 3, integrative modelling (quadrant 4) refers to activities that attempt to predict as-yet unseen outcomes in terms of causal relationships. More specifically, whereas quadrant 3 concerns itself with data that are out of sample, but still from the same (statistical) distribution, here the focus is on generalizing 'out of distribution' to a situation that might change either naturally, owing to some factor out of our control, or because of some intentional intervention such as an experiment or change in policy. This category includes distributional changes for settings that we have observed before (that is, setting an input feature to a specific value, rather than simply observing it to be at that value) as well as the more extreme case of entirely new situations (that is, setting an input feature to an entirely new value that we have never seen before). Integrative modelling therefore requires attention to quadrant 2 concerns about estimating causal, rather than simply associational, effects[49], while simultaneously considering the impact of all such effects to forecast outcomes as accurately as possible (that is, quadrant 3). Ideally work in this quadrant would generate high-quality predictions about future outcomes in a (potentially) changing world. However, forcing one's explanations to make predictions can reveal that they explain less than one would like[15,50], thereby motivating and guiding the search for more complete explanations[51]. Alternatively, such a search may reveal the presence of a fundamental limit to predictive accuracy that results from the presence of system complexity or intrinsic randomness[52], in which case the conclusion may be that we can explain less than we would like, even in principle[53].

Perspective

In addition to clarifying the distinction in practice between predictive and explanatory research activities, Table 1 illustrates our second main point: that whereas quadrants 1, 2, and 3 are all amply populated both with traditional and computational social science research, quadrant 4 is—with a handful of possible exceptions that we discuss in detail below—relatively empty. To an extent, the sparsity of quadrant 4 is not surprising. Models that carefully synthesize the causal relationships between different relevant factors to make high-quality predictions of future outcomes are inherently more difficult to formulate and evaluate than models that aim only for explanatory or predictive power in isolation. Nonetheless, we also believe that quadrant 4 activities are rare because they require one to embrace epistemic values that have historically been regarded as standing in opposition to one another; that is, that explanatory insight necessarily comes at the cost of predictive accuracy and vice versa. If this is true, then viewing them instead as complements, wherein each can reinforce the other, repositions quadrant 4 not as a painful tradeoff but rather as an exciting opportunity for new and impactful research.

To be clear, the opportunity highlighted by Table 1 is not that researchers, computational or otherwise, should focus only or even mostly on quadrant 4. To the contrary, an enormous amount of interesting, high-quality social science exists in the other quadrants, and we see no reason for that not to continue. Indeed, even if one's goal is to end up in quadrant 4, it is arguably impossible to get there without spending a good deal of time in quadrants 1, 2, and 3. Nonetheless, as we will argue in the next section, quadrant 4 research activities that explicitly integrate explanatory and predictive thinking are likely to add value over and above what can be achieved in quadrants 1–3 alone; thus quadrant 4 deserves more attention than it has received so far.

Suggestions

The opportunity that we have just highlighted in turn provokes three related suggestions for methodological innovation in computational social science. First, we make our call for the integration of explanatory and predictive modelling more concrete by sketching out some specific approaches to quadrant 4 research. Second, we advocate for an explicit labelling system that can be used to more clearly characterize individual research contributions, identifying both the quadrant to which it belongs and the level of granularity offered by it. Third, we note that open science practices that have been developed within the explanatory modelling community can be adapted to benefit the predictive modelling community, and vice versa.

Integrate modelling approaches

Our first suggestion is to encourage more work in quadrant 4 by identifying concrete ways of integrating predictive and explanatory modelling. At the highest level, simply thinking explicitly about which quadrants our current models sit in can motivate integrative research designs. Take the example of understanding how information spreads through a social network, a question that has received a great deal of attention with the recent availability of data from online social networks that makes it possible to track with high fidelity how content spreads from one person to the next. At this point there have been hundreds, if not thousands, of studies that explore this question[54]. Some sit squarely in quadrant 1, as purely descriptive studies that measure the size and structure of large and representative sets of online information cascades[48,55]. These efforts have provided insights into how content spreads, some of which align with ideas put forth several decades ago[56] and others that challenge them[57].

Other studies lie in quadrants 2 and 3. For instance, there is work in quadrant 2 that aims to identify features of online content that have a causal effect on the spread of information[58]. Here regression models are used to estimate the extent to which a handful of high-level sentiment features (for example, awe, anger, sadness) affect how far content spreads. This work proposes a theory in which content that reflects positive sentiments spreads further than negative content. Conversely, in quadrant 3 is research that uses as much information as possible to passively forecast content popularity[48,59,60]. Here machine learning techniques are used with an eye towards maximizing predictive accuracy, resulting in statistical models that exploit many features without necessarily focusing on which of these relationships are causal as opposed to merely correlational.

As yet, little, if any, work on this problem would fall in quadrant 4; however, such studies are easy to imagine. For example, one might attempt to explicitly predict the spread of content that has been experimentally manipulated, say by changing content that an individual plans to post to affect its emotional valence or by studying how the same piece of content spreads when exogenously seeded to different individuals. Experiments of this sort would immediately reinforce or challenge results from the other quadrants and would also help to formulate predictively accurate causal explanations.

Orienting our attention towards integrative modelling can also inspire new ways of evaluating the robustness of our findings in other quadrants. Specifically, we can ask how well our estimates and predictions generalize under the types of interventions or changes considered in quadrant 4. In practice, this would mean more cross-domain or out-of-distribution model testing: how well does a causal estimate made in one domain transfer to another domain, or how well does a predictive model fit to one data distribution generalize to another? While informal acknowledgements are often made regarding limitations to generalizability, it is currently rare to see explicit tests of this type in published research. Many of our models are likely to fail at these tasks, but it would be better to clearly recognize and quantify the progress yet to be made than to lose sight of developing high-quality, integrative models that would succeed at them.

Methods from one quadrant can also be leveraged to benefit work in another. In quadrant 2 there are recent examples of using methods from machine learning to improve the causal estimates made with existing explanatory techniques, such as matching and instrumental variables[61], as well as to develop new techniques such as adaptive experimentation to more efficiently learn the effects of deploying different policies[62] and 'causal tree' models for estimating heterogeneous treatment effects[63]. Predictive models have also been used here as a benchmark to assess the 'completeness' of explanatory models[51]. Conversely, in quadrant 3 there are prominent examples in which structural causal models have been leveraged to improve the generalizability of predictive models[49,64].

We can also imagine methods that truly sit in quadrant 4. For example, structural modelling in economics and marketing aspires to "identify mechanisms that determine outcomes and are designed to analyse counterfactual policies, quantifying impacts on specific outcomes as well as effects in the short and longer run."[65] An example entails using estimated models of consumer preferences derived from historical choice data to analyse the effect of a proposed merger. While it is rare to find studies that directly assess the predictive power of such models, as they often concern not-yet-implemented changes, such an extension is clearly possible. For example, Athey et al.[66] used data from sealed-bid auctions to estimate bidder values and make predictions about open ascending auctions, and the predictions were then compared to outcomes in those auctions.

Another method that we believe is particularly promising for making progress in quadrant 4 is akin to a 'coordinate ascent' algorithm, wherein researchers iteratively alternate between predictive and explanatory modelling. Agrawal et al.[12] provide an example of this kind of approach, combining the methods of psychology and machine learning. Their starting point was the Moral Machine dataset, a large-scale experiment that collected tens of millions of judgments from participants all over the world solving 'trolley car' moral reasoning problems[67]. The original study was focused on estimating causal effects, manipulating variables related to the identity of the members of different groups who

Table 2 | A label scheme for clarifying the nature and granularity of research contributions according to the four quadrants discussed above

Granularity	Quadrant 1	Quadrant 2	Quadrant 3	Quadrant 4
	Describes something	Tests a causal claim	Tests a (passive) predictive claim	Tests a claim both for causality and predictive accuracy
Low	Reports stylized facts	Tests for a non-zero effect	Predicts directional or aggregate outcomes	Predicts directional or aggregate outcomes under changes or interventions
Medium	Reports population averages	Tests for a directional effect	Predicts magnitude and direction of aggregate outcomes	Predicts magnitude and direction of aggregate outcomes under changes or interventions
High	Reports individual outcomes	Estimates the magnitude and direction of an effect	Predicts magnitude and direction of individual outcomes	Predicts magnitude and direction of individual outcomes under changes or interventions

The rows distinguish between different levels of granularity in each quadrant. By 'directional', we mean results that report only whether a given association or effect is positive or negative in sign, whereas by 'magnitude and direction' we mean not only the sign of a relationship but also the numerical size of the correlation or effect.

could be hit by an out-of-control vehicle and measuring the changes in participants' judgements of the moral acceptability of different outcomes. Agrawal et al.[12] used this dataset as the basis for building a predictive model, using a black box machine learning method (an artificial neural network) to predict people's decisions. This predictive model was used to critique a more traditional cognitive model and to identify potential causal factors that might have influenced people's decisions. The cognitive model was then evaluated in a new round of experiments that tested its predictions about the consequences of manipulating those causal factors.

Clearly label contributions

Our second suggestion is deceptively simple: researchers should clearly label their research activities according to the type of contributions they make. Simply adding labels to published research sounds trivial, but checklists[68], badges[69], and other labelling schemes are already a central component of efforts to improve the transparency, openness, and reproducibility of science[70]. Inspired by these efforts, we argue that encouraging researchers to clearly identify the nature of their contribution would be clarifying both for ourselves and for others, and propose the labelling scheme in Table 2 for this purpose. We anticipate that many other labelling schemes could be proposed, each of which would have advantages and disadvantages. At a minimum, however, we advocate for a scheme that satisfies two very general properties: first, it should differentiate as cleanly as possible between contributions in the four quadrants of Table 1; and second, within each quadrant it should identify the level of granularity (for example, high, medium or low) that is exhibited by the result.

Focusing first on the columns of Table 2, we recognize that the boundaries of the quadrants will, in reality, be blurry, and that individual papers will sometimes comprise a blend of contributions across quadrants or granularity levels; however, we believe that surfacing these ambiguities and making them explicit would itself be a useful exercise. If, for example, it is unclear whether a particular claim is merely descriptive (for example, there exists a difference in outcome variable y between two groups A and B) or is intended as a causal claim (for example, that the difference exists because A and B differ on some other variable x), requiring us to attest that our model tests a causal claim in order to place it in quadrant 2 should cause us to reflect on our choice of language and possibly to clarify it. Such a clarification would also help to avoid confusion that can arise from any given research method falling into more than one quadrant, depending on the objectives of the researcher (see example in Box 1).

Focusing next on the rows, Table 2 is also intended to clarify that it is possible to engage in activities that reveal widely different amounts of information while remaining within a given quadrant. In quadrant 1, for example, a description that specifies the association between individual-level attributes and outcomes tells us more about a phenomenon than one that does the same things at the level of population averages or 'stylized facts' (that is, the sort of qualitative statements that are often used in summaries of scientific work, such as "income rises with education"). In quadrant 2, estimating the magnitude of an effect is more informative than determining only its sign (positive or negative), which is in turn more informative than simply establishing that it is unlikely to be zero. Likewise, estimates of effect sizes made across a range of conditions are more informative than those that are made for only one set of conditions (for example, the particular settings chosen for a lab experiment[14]). In quadrant 3, predictions about outcomes can also be subjected to tests at widely different levels, depending on numerous, often benign-seeming, details of the test[36]. For example: (a) predictions about distributional properties (for example, population averages) are less informative than predictions of individual outcomes; (b) predictions about which 'bucket' an observation falls into (for example, above or below some threshold, as in most classification tasks) tell us less than predictions of specific outcome values (as in regression); (c) ex-ante predictions made immediately before an event are less difficult than those made far in advance; and (d) predictions that are evaluated against poor or inappropriate baseline models—or where a baseline is absent—are less informative than those that are compared against a strong baseline[35]. The same distinctions apply to quadrant 4, with the key difference being that claims made in this quadrant are evaluated under some change in the data-generating process, whether through intentional experimentation or changes that result from other external factors. Requiring researchers to state explicitly the level of granularity at which a particular claim is made will, we hope, lead to more accurate interpretations of our findings.

Standardize open science practices

Our third suggestion is to standardize open science practices between those engaged in predictive and explanatory modelling. Over the last several years, scientists working in each tradition have promoted best practices to facilitate transparent, reproducible, and cumulative science; specifically, pre-registration in the explanatory modelling community[71], and the common task framework in the predictive modelling community[72]. Here we highlight how each community can learn from and leverage best practices developed in the other.

Pre-registration. Pre-registration is the act of publicly declaring one's plans for how any given research activity will be done before it is actually carried out and is designed with a simple goal in mind: to make it easier for readers and reviewers to tell the difference between planned and unplanned analyses. This procedure can help to calibrate expectations about the reliability of reported findings and, in turn, reduce the incidence of unreliable, false-positive results in research that tests a given hypothesis or prediction[27,71]. Specifically, pre-registration reduces the risk of making undisclosed post hoc, data-dependent decisions (for example, which of many possible statistical tests to run) that can lead to non-replicable findings.

Box 1

How to label a contribution

A regression model of the form $\hat{y} = \hat{\beta} x$ can equally appear in all four quadrants, depending on how the equation is applied and interpreted. In quadrant 1, the association between the outcome and the predictor(s) x is simply described without any causal interpretation or claim about predictive accuracy. In quadrant 2, the same model can be estimated but the focus is on the sign, statistical significance, and sometimes size of the estimated coefficient $\hat{\beta}$, often tied to a causal interpretation derived from substantive theory. In quadrant 3, the same equation can again be estimated, but now the focus is on measuring the error (for example, R^2) associated with predicted values of $\hat{y}$ by comparing them with previously unseen observations[45]. Finally, the same model could fall into quadrant 4 if the goal is to compare the predictive accuracy of different theories[51], and potentially to guide the development of new theories[12,84] that are either more predictively accurate or that generalize to a broader set of circumstances.

Until now, pre-registration has been applied almost entirely in the context of what we call explanatory modelling (quadrant 2), where small sample sizes (for example, in randomized controlled trials) combined with undisclosed flexibility in the data analysis and modelling process led to a high incidence of researchers being unable to replicate published results. However, we believe that it could also be valuable for predictive modelling (quadrant 3) where, in spite of much larger sample sizes, researchers still have many degrees of freedom[73] in their analytical choices. Furthermore, pre-registration can offer a cleaner delineation between the data used to train and validate a model (also known as postdiction[71]) compared to the data used to test it (prediction). The former should be used to develop a model, while the latter should be used only once, at the point when all aspects of a model (including its complexity, hyperparameters, and so on) have been determined and it is ready to be evaluated. While this distinction is clear in theory, in practice research can suffer from confusion about validation versus test datasets, or from multiple uses of test sets within the modelling process[74,75].

In practice, pre-registration suffers from a number of limitations that reduce its value and complicate the interpretation of pre-registered findings[71]. On its own, in other words, it is not a panacea. Nonetheless, the increased use of pre-registration in both explanatory and predictive modelling activities would be likely to reduce the incidence of unreliable results and to improve the transparency and replicability of scientific workflows. Reinforcing pre-registration is the related practice of registered reports[76,77], wherein researchers submit their pre-registered research and analysis plan for peer review before carrying out the study. While registered reports also have their implementation challenges, their adoption would place more emphasis on the quality of the questions being asked and the methods used to answer them than on the answers themselves.

Common task framework. A second practice that could be standardized across communities is use of the common task framework[72] to centralize the collective efforts of many researchers in a given field. In this paradigm there is agreement upon a question of interest, a dataset that pertains to it, and a specific modelling task to be undertaken with that dataset to address the motivating question. An organizer then makes some of the data available to participants and declares the criteria by which research efforts will be evaluated. Participating researchers can then iterate between developing their models and submitting them for

evaluation. Importantly, this evaluation happens on a separate, hidden test set that is accessible to the organizer but not the participants, which helps to guard against overfitting to a particular subset of the data.

The common task framework originated in the predictive modelling community where it is often used for 'prediction contests' such as the prominent Netflix Prize Challenge[78]. However, the common task framework has benefits beyond simply increasing predictive performance, and both the predictive and explanatory modelling communities could benefit from adopting it more broadly. In terms of predictive modelling, increased use of the common task framework would result in easier comparison and synthesis between what are currently disparate research efforts. Recalling the task of predicting how information spreads discussed earlier, there are currently many such efforts that are quite difficult to compare because although they claim to tackle the same problem, they each use different datasets, define different modelling tasks, or use different metrics to quantify success[36]. Centralizing these efforts under the common task framework would force a diverse set of researchers to find common ground in deciding on what the real problems of interest are. It would also standardize the evaluation of progress and make it easy to combine insights across studies.

Likewise, the common task framework could be useful for explanatory modelling. In fact, the common task framework can be thought of as a way of scaling up pre-registration and registered reports from individual researchers to collections of research teams or even entire fields. One example is the recent Fragile Families Challenge[50], which tasked researchers with the problem of forecasting different life outcomes for disadvantaged children and families. This use of the common task framework not only centralized efforts on a prediction problem that is important in its own right, but also generated novel questions about the predictability of different life outcomes for the social science community. Another example is the Universal Causal Evaluation Engine[80], which facilitates collective progress on causal inference through the common task framework[79]. The organizers create synthetic data (for which they know the true causal effects) and make it available to participants who can submit estimates of those effects using their preferred methods. This procedure allows for unbiased evaluation of different inference methods across a range of researchers and research problems.

Outlook

The goal of this Perspective is to advocate for advancing research in the computational and social sciences by integrating predictive and explanatory approaches to scientific inquiry. Our suggestions for doing so, discussed in detail above and summarized in Box 2, are intended to clarify existing styles of work as well as providing useful and actionable advice for researchers interested in integrative modelling. At the same time, we note that the suggestions that we make here are not exhaustive, comprehensive, or without challenges: integrative modelling as we have described it is, on its own, neither necessary nor sufficient for our collective success as a field.

Notably, the issue of model interpretability is missing from the framework and suggestions presented above. Specifically, in discussing explanatory modelling, we have focused on the estimation of causal effects, regardless of whether those effects are explicitly tied to theoretically motivated mechanisms that are interpretable as "the cogs and wheels of the causal process"[7]. This is not because we do not find value in uncovering and understanding causal mechanisms, but rather because it is our view that interpretability is logically independent of both the causal and predictive properties of a model. That is, in principle a model can accurately predict outcomes under interventions or previously unseen circumstances (out of distribution), thereby demonstrating that it captures the relevant causal relationships, and still be resistant to human intuition (for example, quantum mechanics in the 1920s). Conversely, a theory can create

Box 2

Summary of suggestions

- Integrate predictive and explanatory modelling
  - Look to sparsely populated quadrants for new research opportunities
  - Test existing methods to see how they generalize under interventions or distributional changes
  - Develop new methods that iterate between predictive and explanatory modelling
- Clearly label contributions according to the quadrant in which they make a claim, and the granularity of that claim
- Standardize open science practices across the social and computer sciences, encouraging, for instance, pre-registration for predictive models and the common task framework for explanatory modelling

the subjective experience of having made sense of many diverse phenomena without being either predictively accurate or demonstrably causal[81] (for example, conspiracy theories).

Interpretable explanations, of course, can be valued for other reasons. For example, interpretability allows scientists to 'mentally simulate' their models, thereby generating plausible hypotheses for subsequent testing. Clearly this ability is helpful to theory development, especially when data are sparse or noisy, which is often the case for social phenomena. Equally important, interpretable models are often easier to communicate and discuss (verbally or in text), thereby increasing the likelihood that others will pay attention to them, use them, or improve upon them. In other words, interpretability is a perfectly valid property to desire of an explanation, and can be very useful pragmatically. It is our opinion, however, that it should be valued on its own merits, not on the grounds that it directly improves the predictive or causal properties of a model.

We also acknowledge that there are costs associated with adopting the integrative modelling practices that we have described. As mentioned earlier, evaluating explanations in terms of their predictive accuracy may reveal that our existing theories explain less than we would like[53]. Likewise, clearly labelling contributions as descriptive, explanatory, predictive and so on may cast our findings in a less flattering light than if they are described in vague or ambiguous language. Pre-registration requires additional time and effort from individual researchers, and some have criticized it as de-emphasizing important exploratory work. Increased adoption of registered reports requires changes to editorial and review processes, and therefore the coordination of many individuals with potentially disparate interests. The common task framework demands a great deal of effort on the part of those organizing an instance of it[82], as well as adoption by others in the field once a task is created. It is also subject to what has been called Goodhardt's law[83]: "When a measure becomes a target, it ceases to be a good measure."

That said, it is our view that wider adoption of these practices would be a net benefit for the field of computational social science. Exploratory work is important and should be encouraged, but pre-registration is crucial in that it helps to distinguish the act of testing models from the process of building them. Registered reports help us to focus on the informativeness of inquiries being conducted without biasing our attention based on the outcomes of those tests. And the common task framework provides a way of uniting sub-fields and disciplines to accelerate collective progress. Most importantly, thinking clearly about the epistemic values of explanation and prediction not only helps us to recognize their distinct contributions but also reveals new ways to integrate them in empirical research. Doing so will, we believe,

facilitate more replicable, more cumulative, and ultimately more useful social science.

1. Watts, D. J. A twenty-first century science. *Nature* **445**, 489 (2007).
2. Lazer, D. et al. Computational social science. *Science* **323**, 721–723 (2009).
3. Salganik, M. J. *Bit by Bit: Social Research in the Digital Age* (Princeton Univ. Press, 2018).
4. Lazer, D. M. J. et al. Computational social science: obstacles and opportunities. *Science* **369**, 1060–1062 (2020).
5. Lazer, D. et al. Meaningful measures of human society in the twenty-first century. *Nature* https://doi.org/10.1038/s41586-021-03660-7 (2021).
6. Wing, J. M. Computational thinking. *Commun. ACM* **49**, 33–35 (2006).
7. Hedström, P. & Ylikoski, P. Causal mechanisms in the social sciences. *Annu. Rev. Sociol.* **36**, 49–67 (2010).
8. Breiman, L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **16**, 199–231 (2001).
   We view our paper as an extension of Brieman's dichotomy (the 'algorithmic' and 'data modelling' cultures), arguing that these approaches should be integrated.
9. Mullainathan, S. & Spiess, J. Machine learning: an applied econometric approach. *J. Econ. Perspect.* **31**, 87–106 (2017).
   This paper explores the relationships between predictive models and causal inference.
10. Molina, M. & Garip, F. Machine learning for sociology. *Annu. Rev. Sociol.* **45**, 27–45 (2019).
11. Shmueli, G. To explain or to predict? *Stat. Sci.* **25**, 289–310 (2010).
    We build on Schmueli's distinction between prediction and explanation and propose a framework for integrating the two approaches.
12. Agrawal, M., Peterson, J. C. & Griffiths, T. L. Scaling up psychology via Scientific Regret Minimization. *Proc. Natl Acad. Sci. USA* **117**, 8825–8835 (2020).
    This paper exemplifies what we call integrative modelling.
13. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
14. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* https://doi.org/10.1017/S0140525X20001685 (2020).
15. Ward, M. D., Greenhill, B. D. & Bakke, K. M. The perils of policy by p-value: predicting civil conflicts. *J. Peace Res.* **47**, 363–375 (2010).
16. Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122 (2017).
17. Watts, D. J. Should social science be more solution-oriented? *Nat. Hum. Behav.* **1**, 0015 (2017).
18. Berkman, E. T. & Wilson, S. M. So useful as a good theory? The practicality crisis in (social) psychological theory. *Perspect. Psychol. Sci.* https://doi.org/10.1177/1745691620969650 (2021).
19. Athey, S. Beyond prediction: Using big data for policy problems. *Science* **355**, 483–485 (2017).
20. Lipton, Z. C. The mythos of model interpretability. *Queue* **16**, 31–57 (2018).
21. Kleinberg, J., Ludwig, J., Mullainathan, S. & Sunstein, C. R. Discrimination in the age of algorithms. *J. Legal Anal.* **10**, 113–174 (2018).
22. Coveney, P. V., Dougherty, E. R. & Highfield, R. R. Big data need big theory too. *Philos. Trans. R. Soc. A* **374**, 20160153 (2016).
23. Gigerenzer, G. Mindless statistics. *J. Socio-Econ.* **33**, 587–606 (2004).
24. Cohen, J. The earth is round (*p* < .05). *Am. Psychol.* **49**, 997–1003 (1994).
25. Bertrand, M. & Mullainathan, S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* **94**, 991–1013 (2004).
26. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
27. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
28. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
29. Meehl, P. E. Why summaries of research on psychological theories are often uninterpretable. *Psychol. Rep.* **66**, 195–244 (1990).
30. Gelman, A. Causality and statistical learning. *Am. J. Sociol.* **117**, 955–966 (2011).
31. Dienes, Z. *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference* (Macmillan, 2008).
32. Schrodt, P. A. Seven deadly sins of contemporary quantitative political analysis. *J. Peace Res.* **51**, 287–300 (2014).
33. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google flu: traps in big data analysis. *Science* **343**, 1203–1205 (2014).
34. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
35. Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M. & Watts, D. J. Predicting consumer behavior with web search. *Proc. Natl Acad. Sci. USA* **107**, 17486–17490 (2010).
36. Hofman, J. M., Sharma, A. & Watts, D. J. Prediction and explanation in social systems. *Science* **355**, 486–488 (2017).
37. Case, A. & Deaton, A. Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. *Proc. Natl Acad. Sci. USA* **112**, 15078–15083 (2015).
38. Oliver, M. L., Shapiro, T. M. & Shapiro, T. *Black Wealth, White Wealth: A New Perspective on Racial Inequality* (Taylor & Francis, 2006).
39. Chetty, R., Hendren, N., Kline, P. & Saez, E. Where is the land of opportunity? The geography of intergenerational mobility in the United States. *Q. J. Econ.* **129**, 1553–1623 (2014).
40. Wagner, C. et al. Measuring algorithmically infused societies. *Nature* https://doi.org/10.1038/s41586-021-03666-1 (2021).
41. Ba, B. A., Knox, D., Mummolo, J. & Rivera, R. The role of officer race and gender in police–civilian interactions in Chicago. *Science* **371**, 696–702 (2021).

Perspective

42. Provost, F. & Fawcett, T. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking* (O'Reilly Media, 2013).

43. Makridakis, S., Wheelwright, S. C. & Hyndman, R. J. *Forecasting Methods and Applications* (Wiley, 1998).

44. Tetlock, P. E. *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton Univ. Press, 2005).

45. Kleinberg, J., Ludwig, J., Mullainathan, S. & Obermeyer, Z. Prediction policy problems. *Am. Econ. Rev.* **105**, 491–495 (2015).

46. Dowding, K. & Miller, C. On prediction in political science. *Eur. J. Polit. Res.* **58**, 1001–1018 (2019).

47. Galesic, M. et al. Human social sensing is an untapped resource for computational social science. *Nature* https://doi.org/10.1038/s41586-021-03649-2 (2021).

48. Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M. & Leskovec, J. Can cascades be predicted? In *WWW '14: Proc. 23rd International Conference on World Wide Web* 925–936 (2014).

49. Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* **62**, 54–60 (2019).
    This paper outlines the need for causal thinking in building predictive models.

50. Salganik, M. J. et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl Acad. Sci. USA* **117**, 8398–8403 (2020).

51. Fudenberg, D., Kleinberg, J., Liang, A. & Mullainathan, S. Measuring the completeness of theories. *SSRN* https://doi.org/10.2139/ssrn.3018785 (2019).

52. Martin, T., Hofman, J. M., Sharma, A., Anderson, A. & Watts, D. J. Exploring limits to prediction in complex social systems. In *WWW '16: Proc 25th International Conference on World Wide Web* 683–694 (2016).

53. Watts, D. J. Common sense and sociological explanations. *Am. J. Sociol.* **120**, 313–351 (2014).
    This paper argues that sociologists should pay more attention to prediction versus interpretability when evaluating their explanations.

54. Zhou, F., Xu, X., Trajcevski, G. & Zhang, K. A survey of information cascade analysis: models, predictions, and recent advances. *ACM Comput. Surv.* **54**, 1–36 (2021).

55. Goel, S., Watts, D. J. & Goldstein, D. G. The structure of online diffusion networks. In *EC '12: Proc. 13th ACM Conference on Electronic Commerce* (2012).

56. Wu, S., Hofman, J. M., Mason, W. A. & Watts, D. J. Who says what to whom on Twitter. In *WWW'11: Proc 20th International Conference on World Wide Web* 705–714 (2011).

57. Goel, S., Anderson, A., Hofman, J. & Watts, D. J. The structural virality of online diffusion. *Manage. Sci.* **62**, 180–196 (2015).

58. Berger, J. & Milkman, K. L. What makes online content viral? *J. Mark. Res.* **49**, 192–205 (2012).

59. Bakshy, E., Hofman, J. M., Mason, W. A. & Watts, D. J. Everyone's an influencer: quantifying influence on Twitter. In *WSDM '11: Proc. Fourth ACM International Conference on Web Search and Data Mining* 65–74 (2011).

60. Tan, C., Lee, L. & Pang, B. The effect of wording on message propagation: topic- and author-controlled natural experiments on Twitter. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics* 175–185 (2014).

61. Liu, T., Ungar, L. & Kording, K. Quantifying causality in data science with quasi-experiments. *Nat. Comput. Sci.* **1**, 24–32 (2021).

62. Hochberg, I. et al. Encouraging physical activity in patients with diabetes through automatic personalized feedback via reinforcement learning improves glycemic control. *Diabetes Care* **39**, e59–e60 (2016).

63. Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proc. Natl Acad. Sci. USA* **113**, 7353–7360 (2016).

64. Charles, D., Chickering, M. & Simard, P. Counterfactual reasoning and learning systems: the example of computational advertising. *J. Mach. Learn. Res.* **14**, 3207–3260 (2013).

65. Low, H. & Meghir, C. The use of structural models in econometrics. *J. Econ. Perspect.* **31**, 33–58 (2017).

66. Athey, S., Levin, J. & Seira, E. Comparing open and sealed bid auctions: evidence from timber auctions*. *Q. J. Econ.* **126**, 207–257 (2011).

67. Awad, E. et al. The Moral Machine experiment. *Nature* **563**, 59–64 (2018).

68. Aczel, B. et al. A consensus-based transparency checklist. *Nat. Hum. Behav.* **4**, 4–6 (2020).

69. Kidwell, M. C. et al. Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PLoS Biol.* **14**, e1002456 (2016).

70. Nosek, B. A. et al. Promoting an open research culture. *Science* **348**, 1422–1425 (2015).

71. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606 (2018).

72. Donoho, D. 50 years of data science. *J. Comput. Graph. Stat.* **26**, 745–766 (2017).

73. Gelman, A. & Loken, E. The statistical crisis in science. *Am. Sci.* **102**, 460 (2014).

74. Rao, R. B., Fung, G. & Rosales, R. On the dangers of cross-validation. An experimental evaluation. In *Proc. 2008 SIAM International Conference on Data Mining* 588–596 (Society for Industrial and Applied Mathematics, 2008).

75. Dwork, C. et al. The reusable holdout: preserving validity in adaptive data analysis. *Science* **349**, 636–638 (2015).

76. Chambers, C. D. Registered reports: a new publishing initiative at Cortex. *Cortex* **49**, 609–610 (2013).

77. Nosek, B. A. & Lakens, D. Registered reports: a method to increase the credibility of published reports. *Soc. Psychol.* **45**, 137–141 (2014).

78. Bennett, J. & Lanning, S. The Netflix Prize. In *Proc. KDD Cup and Workshop 2007* (2007).

79. Dorie, V., Hill, J., Shalit, U., Scott, M. & Cervone, D. Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *SSO Schweiz. Monatsschr. Zahnheilkd.* **34**, 43–68 (2019).

80. Lin, A., Merchant, A., Sarkar, S. K. & D'Amour, A. Universal causal evaluation engine: an API for empirically evaluating causal inference models. in *Proc. Machine Learning Research* (eds Le, T. D. et al.) Vol. 104, 50–58 (PMLR, 2019).

81. Craver, C. F. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience* (Clarendon, 2007).

82. Salganik, M. J., Lundberg, I., Kindel, A. T. & McLanahan, S. Introduction to the special collection on the Fragile Families Challenge. *Socius* https://doi.org/10.1177/2378023119871580 (2019).

83. Strathern, M. 'Improving ratings': audit in the British university system. *Eur. Rev.* **5**, 305–321 (1997).

84. Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D. & Griffiths, T. L. Using large-scale experiments and machine learning to discover new theories of human decision-making. *Science* **372**, 1209–1214 (2021).