# Value judgements and distributional conflict in the context of empirical policy research and artificial intelligence

*Decision making based on data - whether by policymakers drawing on empirical research, or by algorithms using machine learning - is becoming ever more widespread. Any time such decisions are made, we need to pick both objective functions and policy options carefully. Data do not allow us to avoid value judgements, and do not relieve us from taking sides in distributional conflicts. This essay introduces a general framework to clarify this point, and then discusses a series of settings in which the choice of objective has far-reaching and maybe unexpected implications.*

Economics has experienced an empirical turn in the last few decades. We have entered an era of big data, machine learning, and artificial intelligence. Experimental methods have greatly increased in importance in both the social and life sciences. And recent efforts at reforming the publication system promise to improve the replicability and credibility of published findings. One might be tempted to conclude that this increased availability of and reliance on quantitative evidence allows us to dispense with the normative judgements of earlier days. I will argue that the opposite is the case. The choice of objective functions and of the set of policies to be considered matters ever more in all of these contexts.

A famous example in debates about the dangers of artificial intelligence (AI) is the hypothetical AI system with the objective of producing as many paperclips as possible. If sufficiently capable, such an AI system might end up annihilating humanity in the pursuit of this objective. Another example is the design of experiments. The majority of experiments in the social and life sciences are designed based on the (implicit) objective of obtaining precise estimates of causal effects. Such experiments randomly assign treatments using fixed probabilities. But if the goal of experiments is instead to inform policy choices, or to help experimental participants, then we want to adaptively move assignment probabilities toward the best performing alternatives. These and other examples will be discussed in detail below.

The discussions in this essay will be based on a **general framework** which combines the insights of statistical decision theory and of social welfare analysis.
*Statistical decision theory* requires us to specify objective functions, assumptions about data generating processes, and policy spaces. This makes explicit the relative contributions of value judgements, evidence, and the set of policy options considered, in deriving policy recommendations based on empirical research.
*Social welfare analysis* requires us to explicitly specify how we evaluate individual welfare and how we aggregate individual welfare to some notion of social welfare. This makes explicit, in particular, the distributional conflicts involved in policymaking, and forces us to take a side in these distributional conflicts when making policy recommendations. I believe that economists promoting "inclusive prosperity" should

consistently take the side of those who are worse off.

Using this general framework, I will discuss what has and has not been achieved by the **"empirical turn" of economics**, in terms of overcoming the ideological preconceptions embodied in traditional, more theoretically oriented economics research; this discussion follows up on some of the issues raised in Economics After Neoliberalism. I will argue that (i) the empirical turn has indeed allowed facts to trump dogmatic preconceptions in a number of domains, but (ii) any (empirical) methodology necessarily constrains the set of questions you can credibly address, and thus the scope of political imagination (i.e., action spaces), and (iii) evidence is no substitute for value judgements (i.e., the choice of objective function); obscuring these judgements in the name of "evidence based policy making" is the very definition of ideology.

The remainder of the essay will then discuss the relative role of objective functions and data in a number of specific domains. In the context of **artificial intelligence**, we will consider the problem of multi-tasking and of objective functions which are too narrow, as well as the problem of algorithmic discrimination and targeting. In the context of the **design of experiments**, we will consider the objective of informing policymakers, and the ethical imperative of helping participants, which lead to designs that are very different from standard recommendations. In the context of possible **reforms of the publication system**, current efforts aim to eliminate selection in the form of p-hacking and publication bias in order to improve replicability and the validity of statistical inference. But if we instead consider the objective of publishing findings which are useful for policymakers, then there is a strong rationale for selectively publishing surprising findings.

## General framework

To structure my subsequent discussion, I will now briefly and informally review the frameworks of statistical decision theory and of social welfare analysis. Both of these have a long tradition, see e.g. Wald, A. (1950). *Statistical decision functions*, and Burk, A. (1938). *A reformulation of certain aspects of welfare economics*, and countless contributions since then. The implications of these frameworks for empirical policy research are not generally appreciated, I believe.

### Decision theory

The framework of decision theory considers a decisionmaker, for instance a policymaker, who has to choose between policy alternatives based on some data. (See for instance Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, chapter 2.) To fix ideas, think of a policymaker

who has to decide whether to raise the minimum wage, based on the available evidence on past minimum wage increases.
The effect of alternative decisions on welfare depends on some unknown state of the world. The relevant state of the world in our example might be the effect of a possible increase of the minimum wage on employment.
Data are useful insofar as their distribution depends on the unknown state. In our example, data on the employment impact of past minimum wage increases are useful if three conditions hold: (i) They allow for the credible identification of the causal effect of these increases, (ii) we can expect that the increase under consideration now would have a similar effect to those of the past, and (iii) the effect can be estimated with sufficient statistical precision.
The goal of decision theory is to come up with a "good" mapping from data to policy decisions. The quality of any mapping will depend on the unknown state of the world. Choosing a mapping involves a tradeoff across different possible states.

So how is this framework relevant for the present essay? The decision theoretic framework forces us to be explicit about several key ingredients informing a policy decision.
We first need to define our objective function. This is the central point where **value judgements** come in. Evidence, by itself, can tell us nothing about the choice of this objective function. Choosing this objective function involves distributional conflicts, as discussed below.
We second need to define a policy space, that is a space of possible actions that we consider. Choosing this space reflects the limits of our imagination of political possibilities, and is a matter of **agenda setting** in the political debate.
We third need to make assumptions about how the data relate to the underlying state of the world. We also need to pick a prior distribution over this state of the world for the Bayesian approach, or a set of possibilities over which worst-case scenarios are considered for the minimax approach. These assumptions reflect our **beliefs about** how **the world** works. Different methodological approaches to evidence rely on such prior assumptions to differing degrees, but there exist no approaches that can fully rid themselves of such assumptions.

Once these ingredients (objective function, action space, assumptions about the data generating process) are specified, the policy choice problem becomes a reasonably straightforward math problem. The data themselves cannot tell us, however, what objective function and action space we should consider, and they cannot be interpreted without prior assumptions about the data generating process.

### Social welfare functions

Decision theory, as we just described it, allows for all kinds of objective functions. Optimal policy theory in economics, and theories of justice in philosophy, tend to

consider a more specific class of objectives, namely social welfare functions. (See for instance Roemer, J. E. (1998). *Theories of distributive justice*, and the review in chapter 2 of my online textbook http://inequalityresearch.net)

Social welfare functions take as their point of departure a set of individuals belonging to a given "society." For each of these individuals, we need to come up with a way of evaluating their welfare, under a given set of policies. And we need to aggregate individual welfare into some notion of social welfare. We can think of this aggregation in terms of "welfare weights" for each individual. Each individual's welfare weight determines how much we care about increasing their welfare relative to that of everybody else; the welfare weights determine for instance how much we care about an additional Dollar for a homeless person versus a millionaire.

As before, this framework forces us to be even more explicit about several ingredients informing a policy decision.
We first need to decide on the set of **individuals who matter**. All current citizens or residents of a given country? What about immigrants or residents of other countries? What about future generations? What about animals?
We second need to decide how to **evaluate individual welfare**. In terms of achieved outcomes? In terms of resources at the disposal of individuals to achieve their objectives? In terms of options effectively at their disposal? Relative to their own preferences or relative to some objective scale?
We third need to decide how to aggregate. What **relative importance** do we assign to the welfare of **different people**? How much do we care about an additional dollar for a poor person versus a rich person? For a sick person versus a healthy person? Do we care more about inequality along dimensions such as race or gender than about other dimensions?

When making policy decisions, evidence again only comes in after we have made these choices, and is no substitute for them. In particular, and crucially for the topic of "inclusive prosperity," we always have to pick a side in distributional struggles, which is reflected in the choice of welfare weights. Leaving decisions to the market is one way of making this choice. If we leave decisions to the market, we implicitly assign weight to people in proportion to their disposable income (rich people count more).

## The empirical turn of economics

The remainder of this essay discusses a variety of topics through the lens of this general framework, starting with the empirical turn of economics. Publications in economics, especially highly cited ones, have shifted considerably from theoretical research to empirical research in the last few decades, across almost all subfields of economics. (See for instance Angrist, J. et al. (2017). *Economic research evolves:*

*Fields and styles.*) This shift has had a major impact on the role of economic research in policy debates. Many proponents have celebrated this shift as a **"credibility revolution."** Much campaigning in favor of **"evidence based policymaking"** is taking place. And many argue that the task of empirical researchers is to just find out **"what works."** How shall we assess these claims in light of our general framework?

First, I believe it is true that the empirical turn has allowed **facts to overcome some received dogmas**, due to researchers' increased reliance on data and decreased reliance on prior assumptions. A showcase example is the effect of minimum wage increases on employment. Traditionally, drawing on the standard competitive model of labor markets, economists firmly believed that increasing the minimum wage reduces employment. A large number of empirical studies (starting with Card, D., and Krueger, A. B. (1993). *Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania.*) have found no evidence of such a negative impact, for the ranges of minimum wage levels observed in the United States in recent decades. As a consequence both the policy consensus and the theoretical debate regarding minimum wages have shifted considerably.

Second, and running somewhat counter to the first effect, I believe that the empirical turn has significantly **constrained the political imagination** in terms of the sets of policies considered. This effect is noticeable for instance in the field of development economics, the subfield of economics that has most enthusiastically embraced randomized field experiments as its leading source of evidence. There is a large space of policies that is not amenable to randomized field experiments. This includes policies affecting many of the classic topics of development, such as "modernization," dependency on a metropolitan core, import substitution, class conflicts in post-colonial nations, etc. In contrast to these older macro-social questions, the policies most often studied in more recent field experiments are individual level (or at most village level) "treatments," such as the provision of deworming pills, bed-nets, or school books.
The example of development economics reflects a more general point: If we restrict our debates to policies whose effects we can "credibly" evaluate, and if we adopt a certain methodological perspective on what credibility means, then we necessarily restrict the space of policies which is up for debate.

Third, **evidence is no substitute for value judgements**, nor is the existence of evidence independent of value judgements. Take for instance the literature on the impact of the expansion of the Earned Income Tax Credit (EITC) in the United States in the 1990s. This expansion has generally been celebrated as a success, since it resulted in some redistribution to the (working) poor, while simultaneously increasing labor supply. But is it actually good to increase labor supply? Closer inspection reveals that this effect is desirable only if we make a value judgement that distinguishes the deserving (working) poor from the undeserving (non-working)

poor. If we instead take a conventional utilitarian perspective, then increased labor supply in the presence of negative marginal taxes generates "dead-weight loss;" an unconditional basic income would be preferable to subsidies of low wage labor (such as the EITC) from the perspective of utilitarianism.

Lastly, there is the related issue of **what evidence is even generated**. Both the collection of survey data or administrative data, and the empirical analysis of them, are challenging tasks. Many questions never get addressed by quantitative empirical research, either as a consequence of passive omission, or as a consequence of active opposition to the generation of evidence. An example of the latter would be the collection of data on wealth inequality in the European Household Finance and Consumption survey, which had to overcome significant political opposition.

To summarize, the "empirical turn" has pushed economics to rely more on data and less on prior assumptions. This shift has allowed the field to overcome some dogmatic beliefs, but also restricted the set of policies most often discussed. Economics is not *necessarily* a tool of reaction. It has potentially much to contribute to a progressive or even radical policy agenda. But economic research requires, even in its empirical flavor, many normative choices. Obscuring or denying these choices in the name of "evidence based policy making" (or the objectivity of "artificial intelligence," to which we will turn next) is the very hallmark of ideology.

## Artificial Intelligence

A common conception of Artificial Intelligence (AI) and Machine Learning (ML) among economists is that these fields primarily concerned with prediction problems, and that they can be useful primarily for policy problems which depend on accurate predictions (see for instance Kleinberg, J. et al. (2015). *Prediction policy problems.*). This might be an accurate description of one sub-field of AI, supervised machine learning.

In general, however, the field of AI takes a considerably broader perspective. One of the leading textbooks on AI (Russell, S. J., and Norvig, P. (2016). *Artificial intelligence: a modern approach*, chapter 2) defines **AI as the construction of rational agents**. Agents receive information from their environment through perceptors (sensors), and act on their environment through actuators. The agent program maps sequences of percepts into actions. A rational agent is then defined as follows.

> For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has.

This general definition covers a wider range of different approaches to the construction of rational agents. What is noteworthy for our purposes is that this definition is quite similar to the framework of statistical decision theory sketched above. In particular, any construction of AI systems requires a careful choice of objective functions, that is, value judgements. We will discuss here two aspects of this choice that appear particularly salient, multitasking, and discrimination and fairness.

### Multitasking

In many ways, the designer of an AI system faces similar issues to a managerialist technocrat who wishes to design incentive pay systems based on quantitative measures of performance, in order to guide the efforts of human agents. A controversially discussed example of the latter is incentive pay for teachers. One way to measure teacher performance is based on the improvements of their students on standardized tests over time. Based on the observation that these improvements appear positively correlated with long-run student outcomes, there is a push to tie teacher pay to these improvements.
Nobel-memorial-prize winning work on multitasking (Holmstrom, B., & Milgrom, P. (1991). *Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design.*) has provided a key argument why this might not be a good idea. Agents generally devote **effort to many tasks**. **Incentivizing only one** (or a subset) of these tasks leads to **effort being diverted** from other tasks. This might lead to a dramatic worsening of overall performance. For teachers, helping students improve their test-taking skills might be one of the tasks they invest effort in. Helping students develop on dimensions not measured by standardized tests, say creativity, curiosity, or oral communication, would fall by the wayside when teachers are induced to teach to the test.
The stakes get more dramatic when we turn to medical contexts. Suppose doctors were to get paid based on the number of patients they treat. Then clearly the quality of care for any given patient would suffer. Or suppose they were to get paid based on the survival or recovery rates of their patients. Then they might end up turning away all but the healthiest patients.

Extreme versions of this line of reasoning have been discussed in the AI community, with regards to fears about the **annihilation of humanity by some super-human artificial intelligence**. A famous example was formulated by Nick Bostrom (2003) (see also Universal Paperclips):

> Suppose we have an AI whose only goal is to make as many paper clips as possible. The AI will realize quickly that it would be much better if there were no humans because humans might decide to switch it off. Because if humans do so, there would be fewer paper clips. Also, human bodies contain a lot of atoms that could be made into paper clips. The

> future that the AI would be trying to gear towards would be one in which there were a lot of paper clips but no humans.

Such scenarios are often dismissed as being unrealistic, given the distance of present day technical capabilities from the possibility of constructing a general AI with such capabilities. Such dismissal misses the point, I believe. Variants of this danger in more mundane settings are very much a reality.
Consider for example the use of machine learning to match unemployed workers to jobs. Such algorithms are being used by employment agencies in order to maximize the rate of proposed matches which result in employment. This objective, combined with penalties for job seekers who do not accept proposed matches, might lead to systematically placing workers in jobs for which they over-qualified, resulting in de-skilling, wage declines, and reductions in worker welfare.

These examples, and many others, drive home the point that we need to very carefully specify and regulate the objective functions of AI systems - even in their mundane, present-day versions which do not yet threaten the annihilation of humanity. Otherwise we might end up like the Phrygian **king Midas**, starving because our wish of turning everything into gold was granted.

### Discrimination and fairness

One of the key uses that machine learning has found thus far is to provide **individually customized treatments**. In online settings this might involve customized search results in search engines, customized ads, and customized feeds in social media, individual-specific pricing in online shops, automated hiring decisions by human resource departments, automated credit approvals based on credit scores, etc. These algorithms aim to make optimal decisions for the **objective of profit maximization**. But how does this objective align with **socially desirable outcomes**? How, in particular, is individually differentiated treatment for profit maximization compatible with notions of fairness?

Algorithms might be criticized on the basis of various, **increasingly stronger, notions of fairness**. (The following argument has previously appeared in the blog entry *The politics of machine learning*.) First, prediction algorithms might systematically get the predictions wrong, reproducing pre-existing biases against certain groups. Algorithms assessing job-candidates, for instance, might discriminate against women, since women in the past were less likely to be promoted on the job. Second, it might be considered discriminatory to use certain variables in making decisions. A judge, for instance, should not treat a black defendant differently from a white defendant with the same biography and criminal history. Third, even if certain variables such as ethnicity or gender are excluded from prediction, algorithms might still treat the corresponding groups systematically differently. With enough other

data, it is fairly easy for the algorithm to "guess" someone's ethnicity or gender, and to implicitly base decisions on this guess. Fourth and finally, it might be argued that fairness demands the same treatment of everybody not only between groups, but also within groups. Why should some people go to jail and others not, based on crimes they have not committed yet? And why should some people pay more for goods they purchase relative to other people, just because they need them more?

These four criticisms are based on increasingly stronger notions of justice or fairness. The development of technology pushes us to question each and move on to the next one: Even if algorithms get predictions right, they will discriminate. Even if explicit discrimination is prohibited, algorithms are able to discriminate implicitly. Even if equality across groups is enforced, within group inequality might be rising.

The notions of fairness just described are framed rather narrowly: Is it justifiable that people in a particular situation are treated differently? It is advisable to take a step back and ask a bigger question: What **impact** does **differentiated treatment** have **on social inequality more broadly**, and how do we evaluate this impact? Asking this question is precisely what a social welfare framework compels us to do. What "big data" and "machine learning" do to economic and social inequality more generally depends on how new inequalities in treatment align with pre-existing inequalities. If the new algorithms are less likely to hire women, more likely to send black defendants to jail, or charge poor people higher prices, then they increase inequality. The opposite might happen as well, if unequal treatment is negatively correlated with pre-existing inequalities. For instance, targeted prices might end up being lower for poorer individuals, because their ability to pay for various goods is smaller. It is hard to say how these dynamics will play out, and most likely they will play out differently in different settings. It is all the more important, then, to stay vigilant and to observe how newly generated inequalities in treatment correlate with old inequalities, and how they therefore impact overall social inequality.

## Experiments for policy choice

We have discussed the role of objective functions in the context of policy choice based on empirical evidence, as well as its automated cousin, autonomous decision making by AI agents. At this point some might concede that values matter for the use of data in policymaking, while the conduct of research itself can remain free of value judgements. I will argue in the following that this is not the case, focusing on the design of experiments, and on the organization of the publication system.

As discussed above, randomized field experiments have quickly become a leading source of evidence for development economics, but also for policy research in rich countries, in domains such as education, medicine, public health, and public finance. The design of the typical randomized controlled trial (RCT) follows classic prescrip-

tions: Compare a treatment to a control. Split your sample equally between the treatment and the control group. If you have baseline covariates, stratify on those, and split your sample equally within each of the strata. After conclusion of the experiment, compare average outcomes in the treatment and control group (possibly adjusting for covariates) to estimate the average treatment effect. Use this estimate to construct a statistical test for the hypothesis that the true effect equals zero.

Are these prescriptions sensible? Well, it depends. Yet again, optimal behavior depends on the choice of objective function. You might care about getting precise estimates, you might care about getting a policy choice right to maximize some notion of welfare, or you might care about doing right by the participants of your experiment. Each of these objectives leads to different prescriptions for the design of experiments.

Suppose your goal is indeed to get a **precise estimate** of the average causal effect of your treatment, relative to the control, where precision is defined in terms of a small squared estimation error. Then the standard design is indeed quite sensible.

But what if your goal is not to get precise estimates, but instead to **choose good policies** based on the outcome of your experiment? This is the problem we consider in (Kasy, M., and Sautmann, A. (2019). *Adaptive Treatment Assignment in Experiments for Policy Choice.*). (Part of the following discussion also appeared in the blog entry *Experiments for policy choice*.) Trying to identify the best policy is different from estimating the precise impact of every individual policy: As long as we can identify the best policy, we do not care about the precise impacts of inferior policies. Splitting the sample at the outset will lead us to "waste" sample size learning about the precise impacts of a treatment that is clearly suboptimal. The key to our proposal is staging: rather than running the experiment all at once, we propose that you should start by running a first round of the experiment with a smaller number of participants. Based on this first round, you will be able to identify which treatments are clearly not likely to be the best. You can then go on to run another round of the experiment where you focus attention on those treatments that performed well in the first round. This way you will end up with a lot more observations to distinguish between the best performing treatments. And if you can run the experiment in several rounds, you can do even better. As we show, such a procedure, where you shift towards better treatments with the right speed, gives you a much better chance of picking the best policy after the experiment.

Both proposals discussed thus far - the standard design, and adaptive designs for policy choice - ignore the wellbeing of experimental participants. Is that ethically acceptable? Not according to Immanuel Kant, who famously posited that you should

> Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end.

This brings us to a third possible objective for experimental design, **maximizing the outcomes of** experimental **participants**. This objective motivates multi-armed bandit algorithms. This leads to recommendations which are different again from those for choosing policies. Consider the approach outlined above, where you run the experiment in two rounds. If you have a bandit objective, you will assign all participants in the second round to the one treatment that performed best in the first round. With continuing experimentation, you won't want to be quite as extreme. It remains the case, though, that the objective of maximizing participant welfare compels us to move more quickly toward the better performing options than you would for the policy choice objective.

In an ongoing field experiment in Jordan, we use such an experimental design geared toward maximizing the outcomes of participants (Caria et al. (2019) *Job Search Assistance for Refugees in Jordan: An Adaptive Field Experiment*.)

## Reforming the publication process

Let us now, lastly, turn to the debates about publication bias, replicability, and the various reform efforts aimed at making empirical research in the social and life sciences more credible. Once again we will see that the desirability of these reforms hinges on value judgements, that is, on the presumed objective of scientific research and publishing. It is far from obvious how to choose the right objective.
A key concern in the debates about replicability is the issue that published findings are selected. This happens, first, because researchers don't write up all their empirical findings when analyzing data; one form of this is sometimes discussed as "p-hacking." This also happens because journals don't publish everything; if they selectively only publish certain findings, this leads to "publication bias."

### Replicability and validity of inference

Why is **selection** of findings for publication, by researchers or by journals, a problem? Because it **makes** all our standard **inferences invalid** (cf. Frankel, A., and Kasy, M. (2019). *Which findings should be published?*). Standard inference methods are valid if and only if publication probabilities do not depend on findings in any way (dependence on standard errors is allowed). Suppose, for instance, that a journal in nutrition only publishes studies which find that some type of food decreases the chance of cancer. Even if food consumption is completely irrelevant for the risk of cancer, by pure statistical chance some studies will find such a connection. And if the journal only publishes such studies, readers are erroneously induced to conclude that there are all kinds of miracle cures.

We do have ample **evidence that publication is selective**, albeit to different

degrees and in different ways across different empirical fields (see for instance Andrews, I., and Kasy, M. (2019). *Identification of and correction for publication bias*). Recognition of these facts has motivated reform initiatives aimed at increasing the replicability and credibility of published research by reducing selection in the publication process. These are valuable initiatives that are likely to improve the standards of scientific evidence. They do raise, however, the question of what it is that reforms of academic research institutions and norms wish to ultimately achieve: What is the objective function of scientific research and publishing?

### Relevance of findings for decision makers

Consider, as an example, clinical research on new therapies. Suppose that in some hypothetical area of medicine a lot of new therapies, say drugs or surgical methods, are tested in clinical studies. Suppose that most of these trials don't work out - the new therapies just don't deliver. Absent a publication of successful clinical research, no doctor would implement these new therapies. And doctors have limited time - they are not going to read hundreds of studies every month. But they might read some.

In this hypothetical scenario, which findings should be published? That is, which subset of studies should doctors read? In order to improve medical practice, it would arguably be best to tell doctors about the small subset of new therapies which were successful in clinical trials. Those are the therapies they should incorporate into their practice. If this is the selection rule used for publication, however, published findings are biased upward. Replications of the published clinical trials will systematically find smaller positive effects or even sometimes negative effects. This reasoning suggests that there is a deep tension between relevance (for decision making) and replicability in the design of optimal publication rules.
In Frankel, A., and Kasy, M. (2019). *Which findings should be published?*, we argue that this type of logic holds more generally, in any setting where published research informs decision makers and there is some cost which prevents us from communicating all the data. In any such setting, it is optimal to selectively **publish surprising findings**.

### A multi-tiered publication system?

These considerations leave us with the practical question of what to do about the publication system. How shall we trade off these conflicting objectives? Can we have validity and relevance at the same time? A possible solution might be based on a **functional differentiation of publication outlets**, which could build on the present landscape, while making the differences of objectives and implied publication

policies across outlets more explicit. Such a differentiation avoids having to sacrifice one of these objectives (e.g. relevance) for the sake of another (e.g., validity and replicability).

There might be a set of top outlets focused on publishing surprising ("relevant") findings, subject to careful quality vetting by referees. These outlets would have the role of communicating relevant findings to attention-constrained readers (researchers and decision makers). A key feature of these outlets would be that their results are biased, by virtue of being selected based on surprisingness.
There might then be another, wider set of outlets that are not supposed to select on findings, but have similar quality-vetting as the top-outlets, thus focusing on validity and replicability. For experimental studies, pre-analysis plans and registered reports (results-blind review) might serve as institutional safe-guards to ensure the absence of selectivity by both researchers and journals. Journals that explicitly invite submission of "null results" might be an important part of this tier of outlets. This wider set of outlets would serve as a repository of available vetted research, and would not be subject to the biases induced by the selectivity of top-outlets.

## Conclusion

We have reviewed the general frameworks of statistical decision theory and social welfare functions, and have argued that any use of data requires value judgements, and in particular requires taking a side in distributional conflicts. Using these frameworks, we have discussed a diverse set of areas in which value judgements (choice of objective function, choice of welfare weights, choice of possible policies to be considered) play a crucial role that is often under-appreciated.

Who should make these value judgements? This is clearly not a task to be left to supposed experts or technocrats. Value judgements and the resolution of distributional conflicts need to be subjected to public debate and a democratic process. That said, just as data cannot provide "objective" answers that are a substitute for such judgements, neither can the outcomes of majority votes provide a substitute for such judgements. Economists promoting "inclusive prosperity" need to be explicit about the necessity of judgements, and they need to partake in a public debate about them, but they should also unapologetically take the side of those worse off when making these judgements.

### Thanks

I thank Pirmin Fessler, Zoe Hitzig, Suresh Naidu, and Dani Rodrik for helpful feedback and comments.