Ten Governance Concerns about The Nature and Use of Data

Mark Amadeus NOTTURNO
Interactivity Foundation

Abstract. The emergence of big data and the data revolution raises a number of governance concerns about the nature and use of data. This chapter describes nine such concerns that an international group of data experts articulated and explored during a series of online discussions devoted to that issue. I then conclude the chapter by arguing against a common interpretation of evidence-based policy decisions – namely, the use of data to try to justify or promote public policy proposals – and in favor of a more critical, and self-critical, approach to evidence-based public policy decisions that uses data to criticize policy proposals instead of trying to justify them. I also argue that we should pay greater attention to the underlying philosophical beliefs, concerns, goals, values, interests, and priorities that motivate them.

Keywords. Big data, data revolution, governance concerns, public policy, quality of data, interpretation of data, ownership of data, intellectual property, privacy of data, transparency of data, use and misuse of data, moral nature of data, effect of data upon data, public and private good of data, evidence-based policy decisions, critical use of data, Interactivity Foundation

Introduction

The Interactivity Foundation (IF) is a private, non-partisan foundation that conducts governance projects on areas of public policy concern with the aim of selecting the concerns about those areas that are most useful for public discussion, and developing contrasting public policy possibilities for addressing them. An IF project usually involves two different panels – each with six to eight members – one consisting of experts and professionals working in the area of concern, and the other of interested citizens. IF usually guarantees the confidentiality of its discussions, and does not reveal the names of the panelists who participate in them, in order to encourage them to speak openly and without fear of retribution from the institutions for which they work. IF's aim in sponsoring these projects is not to advocate any of the policy possibilities that panelists develop, let alone to "fix" public policy, but to stimulate public discussion about the governance concerns that people have and the different policy possibilities for addressing them.

I have been an IF Fellow for the past thirteen years, and have conducted projects on privacy; science; property; democratic nation building; money, credit, and debt; global responsibility for children; and the future of employment. I also use a video conference platform to conduct weekly online "public discussions" of the policy possibilities in IF's reports.

A former panelist who works at an international organization recently asked if I could facilitate similar discussions about the nature and use of data. His motive was clear. Today our electronic information technologies have ushered in the age of "big data" and we are experiencing a data and information explosion that outstrips anything seen in the history of mankind, enabling us to collect and sort through more data in just the past two years than in all the time that went before. The opportunities for innovation that these technologies make possible may be limitless. And The United Nations and partner institutions, such as the World Bank and the Department for International Development, have recently called for a “data revolution” to support the design and implementation of the new global development agenda called “The Post-2015 Sustainable Development Goals”.

Such a bold undertaking should clearly be grounded in a realistic sense of the limitations of data and the improvements that are needed in its governance. So I agreed to conduct a series of discussions to explore the governance concerns that people might have about the nature and use of big data, and public policy possibilities for addressing them. These discussions are not part of a regular IF project. All of their participants, on the contrary, are professionals working in the area of data collection and design. But I have conducted them as if they were, since, well, that is what I do. A group of twelve or so international data experts has been meeting for two-hour discussion sessions on a weekly basis for several months now. These people hold, or have held, data research, engineering, and policy-making positions in government, civil society, and academia. Some of them work, or have worked, in data development for large international organizations. Some teach, or have taught, about data at colleges and universities. Some have founded commercial data start-ups. And many of them have worked in worldwide projects.

In what follows, I will describe nine governance concerns about the nature and use of data that they have articulated and explored during our discussions. These concerns, unless otherwise indicated, do not necessarily reflect the consensus of the group, but only the range of its concerns. In the last section of the chapter, I will discuss a special concern of my own, namely, the use of data in making so-called “evidence-based” public policy decisions.

1. The Quality of Data

Regardless of the quantity of data we collect, there is always a question about its quality: whether and to what extent it is accurate, whether and to what extent it is credible, whether and to what extent we can rely upon it. The quality of data is of public concern because people very often use data to try to justify public policy decisions. I will return to this issue later in this chapter, where I will argue that we should not use data in this way. But here, the point to be made is that data itself is never self-defining. It may be incomplete, unreliable, or even mistaken. There are sampling errors, selection bias, confirmation bias, and a gap between the raw data and our interpretations of it. People and institutions may thus look for and collect data about certain concerns and not about others. And they may disagree about what is and is not a concern. They may select data that confirms their pre-conceived beliefs and furthers their policy interests, aims, and agendas – or interpret the data they collect in ways that do so – and entirely ignore data and interpretations that do not confirm their beliefs, or run contrary to their policy interests, aims, and agendas. So it is, perhaps, not too surprising that the quality of data was the single greatest concern that our discussion participants had about big data.

Here, a lot depends upon what kind of data we are looking for, what questions we ask, and who is asking and answering them. Our discussion participants said that this is the crux of the issue; that there is no data without human intervention, perception and the creation of meaning, and the quality of the data we collect will ultimately depend upon who has the power to ask the questions. The structure of social relations is thus at the heart of the issue, since data itself has no existence outside the realm of social relations, and the questions we ask, how we ask them, who we ask, who does the asking, and who answers ultimately all reflect what the people doing the research regard as important. The participants agreed that big data might open a realm of unlimited possibilities. But they also agreed that a big data revolution might be counter-productive if it leaves us with big misinformation or, worse still, big disinformation.

The participants generally felt that the accuracy and reliability of the data that we collect and use is far more important than its quantity, that data can be collected in ways that are biased, both consciously and unconsciously, and that numbers can be deceptive, especially if the people who use them do not understand what they mean. They thus spoke about the quality of data in terms of its *integrity*. The integrity of data is defined, at least in part, by how we collect it; and the data we get from some sources may be incomplete, inaccurate, and unreliable. They said that using data is always a gamble, that people make their bets depending upon what it says, and that they should be able to choose between low risk, low return data and high risk, high return data. They also said that it is probably impossible to eliminate bias entirely, but that public entities, such as governments, should generally use data only if it has been collected, analyzed, and interpreted by relatively unbiased sources – and only if it has been vetted by relatively unbiased people whose concerns, beliefs, values, goals, and interests are diverse enough to reduce bias. And they said that everyone who uses data should be clear about the nature of the raw data, about what questions were asked and what methods were used to collect it, about how the people who evaluated it got from the raw data to their interpretations of it, about the limits of the data itself, and about any doubts that might exist about its accuracy and reliability.

The participants worried that detailed information about how complete, accurate, and reliable the data is – let alone how it was collected, evaluated, and selected – often does not travel with a dataset, so that it is difficult for users to determine for themselves whether and to what extent it is accurate and unbiased, let alone reliable. They said that we currently do not have any good way to measure the quality of the data that we use, or even how complete it has to be in order to be useful, and that we should find a way to measure the quality of data before we put too much faith in it. They thus felt the need to find ways to evaluate and measure the quality of the data we use in order to feel comfortable about it. But they also worried about how we can know whether and to what extent data is accurate, reliable, and complete – especially now that we have so much of it to analyze and so few people who are qualified to evaluate it. They said that it is simply not clear whether having more data increases the accuracy of evidence, or whether it simply increases the level of noise in the decision making system. And they worried about whether we were now focusing too much upon data itself – because of our increased abilities to collect it and our interest in the technologies we use to collect it – and that this focus might crowd out or obstruct our focus upon how we collect it, how we interpret it, what we are trying to do with it, and other governance concerns we might have about it.

2. The Interpretation of Data

Our discussion participants also drew a sharp distinction between the data we collect and how we interpret it – which, some of them said, is what transforms data into information. The bits of data we collect are not themselves information. Information is what we derive from them. The bits themselves are meaningless unless or until we interpret them. And we cannot interpret them, or derive anything from them, without a theory or hypothesis. We must thus make inferences in order to make sense of data. Drawing inferences from data is the way we interpret it and transform it into information and ultimately make productive use of it. And the meaning, or information, we derive will ultimately depend upon the questions we are trying to answer. All of this underscores the fact that data does not simply "exist" out there; that it is, on the contrary, a human creation that is mediated by the methods and tools we use to collect it, and by the hypotheses and theories we use to interpret it,and that its very existence depends upon and emerges from the interaction of humans with other humans and their environments. The participants said that questions always shape their answers, and that the fact that we need to impose a theory or hypothesis upon data in order to interpret it and derive meaning from it can very easily prejudice it. They said that a lot here will depend upon what we want to do with the data; that we sometimes try to measure things that are not so easily quantifiable, such as poverty, the quality of life, and discrimination, and that there always are, or should always be, questions about what the data actually measures and means.

All of this points to the epistemological side of big data.

Here, one might think that the theories and hypotheses we use to interpret data should be suggested by the data itself. This may happen if we are scrubbing the Internet with no theory or hypothesis in mind and notice certain correlations in the data we collect. Indeed, the very ability to scrub the Internet is, perhaps, the reason why some people have suggested that big data means the end of theory. Theories are important because they allow us to make inferences about things we have not observed and may not be able to observe. And the fact that we can now collect so much data may suggest that we no longer need them. It is, however, usually the other way around. And it is easy to see why. If we ask people to simply collect data, or to look for the relevant data, they probably would not know what we mean, let alone what to do. For data can quite literally be anything. So they will want to know what kind of data we are looking for. Are we looking for data about the population or data about the economy or data about the weather? About literacy rates in third world countries, or social diseases in New York City? Are we looking for data about illegal immigration, gross domestic products, agricultural outputs, or tax cheats? The possibilities, of course, are limitless. But the specification of what to look for implicitly places limits upon what we do look for and, of course, upon what we find.

One can see something like this in the fact that we typically define the success metrics for a project at the very beginning of a project. Such success metrics, in turn, define the kind of data we should regard as indicating the success or failure of a project. But while the success metrics tell us what we should look for and measure, they are typically proxies for the real indicators of success, which may be far more difficult or even impossible to measure directly. Here, our discussion participants said that focusing upon proxy success indicators and collecting data pertaining to them is important for justifying funding decisions and appeasing donors. But they also said that focusing too exclusively upon proxy success indicators may lead us to ignore other indicators, or give short shrift to other measures, or forget that they are only proxies for the real indicators, or ignore the gap between the proxy success indicators and the real indicators altogether; that this, in the end, may lead to bad decision-making, and that success metrics and indicators should be constantly reevaluated as a project proceeds.

The participants said that the data we look for will typically be determined by what we want to do with the data we find – so all data, like all observation, will necessarily be both theory and value laden, and in some sense determined by our goals. And even if we do scrub the Internet, what we will end up seeing in the data we collect will ultimately be, consciously or unconsciously, theory and value laden as well. They said that, regardless of whether or not big data means the end of theory, it might easily leave us with more noise than useful information, with more data than we can usefully analyze, with hypotheses that may be false or even meaningless, or with hypotheses that we do not have the time or interest to test. This is especially true when it comes to big data, since big data makes it easier to commingle diverse datasets that we might not otherwise be able to compare with each other. The ability to compare diverse datasets is, no doubt, one of the great potentials of big data. But making sense of the various correlations that might arise from the aggregation of multiple and seemingly unrelated datasets is also one of the greatest challenges that it poses for us. High degrees of correlation may suggest, and may lead us to discover, causal relationships at play in places where we might never otherwise expect to find them. But big data does not change the fundamental epistemological realities, and we should never forget that correlation by itself never implies causality, no matter how prevalent or perfect it may seem to be, and that the fact that people often derive causal relationships from data about correlations only means that they are, consciously or unconsciously, imposing a theory on it.

3. The Use and Misuse of Data

The discussion participants also voiced a number of concerns about the use of data, and especially its misuse: not only for criminal purposes, but also for discrediting or disgracing a data owner or subject, thereby bringing them into disrepute. They said that this latter concern has led data owners and subjects to try to put controls on how data can be used downstream, such as Canada's infamous first Open Data License, which prohibited data from being used to bring the Government of Canada into disrepute, and said that the Government of Canada would be the sole judge of whether or not it does so. They said that the Open Data movement normally regards criminal activity pertaining to data as primarily a matter for the police to worry about, rather than the data owners and subjects, but that there have been cases in the UK in which certain anti-fraud procedures simply assumed that certain sensitive data would never be made public and had neglected to tell its owners and subjects that they were relying upon keeping it secret.

The participants were also concerned about the use of data to persuade policy makers to make certain kinds of policy decisions, and about how to move the evidence that data might represent up on their list of priorities. They said that policy is ultimately about politics and all about power, but that there are always different ways to interpret and use the data we collect. Certain powerful institutions – such as the United Nations, the International Monetary Fund, the World Bank, and commercial corporations – may claim to be neutral, and may indeed want to be neutral and strive to be neutral, but ultimately they are not neutral and, indeed, cannot be neutral in this respect, since their very creation was based upon the pursuit of certain philosophical and political-economic ideas and goals. Such institutions have thus articulated the problems they want to solve, and they have emphasized the importance of evidence and evidenced-based solutions. But they need data to test the progress they are making in solving those problems, and programmatic measures to test the efficacy of different solutions. Here, the participants said that you cannot expect the same from an organization that is blind as from an organization that can see, and that there may well be a responsibility *not* to be neutral that comes with power. For if an organization has the power to gather and analyze data, then the question arises whether it can or should remain neutral if it discerns patterns in the data that might allow it to help people. But they also said that we would probably see leaders of such organizations criticized more and more often for not using data appropriately.

There was also a sense that some individuals and institutions may put too great an emphasis upon using data, especially when they have the responsibility for accounting for the use of public money, and that an obsession with using data and metrics may easily crowd out discussion about its quality, what it means, and the best ways to use it. The participants said that there is a danger that people may expect too much from large data sets and advances in information processing, and that things can get very complicated when institutions use data in an attempt to justify to their shareholders and donors that their projects work. They said that we do not, or should not, collect data for its own sake; that we should collect it in order to solve problems. And that it is not at all clear whether the fact that we have new technologies that enable us to collect more and more data means that we will get better and better data, or that we will be better equipped to solve our problems. Indeed, some of the participants seemed to think that we might be better off collecting *less* data, especially if we could ensure that the data we collect and use is accurate and reliable.

Our participants also said that the concerns that citizens have about the use of data might be different depending upon whether the data is being used by governments, international organizations, or private agents. Some citizens are thus concerned about their governments using data for the purpose of surveillance and discrimination, while others are pressuring their governments to make data more open to improve their own transparency and accountability. Citizens have yet to voice strong concerns about the use of data by international organizations, which generally use data primarily to market their programs by demonstrating their progress and effectiveness. But some citizens are concerned about uses of data by the private sector that are not consistent with their preferences, and others are pressuring private companies to make more user data open so users can have more control over the information that is being developed by them. Be this as it may, our participants said that there is a growing concern among citizens that data can be passed on and used by parties who have no knowledge about the context in which, and the purposes for which, it was collected – and that the potentially nefarious uses of information drawn through the combination of diverse data sets is what concerns them most. They said that this is the greatest concern that citizens have about the use of data, that it is a concern that has yet to be clearly addressed, and that we should develop a clear code of ethics to determine which uses of data are appropriate and which are not.

4. The Ownership of Data

The discussion participants also raised intellectual property concerns about who owns the data that is collected. Published information is often subject to copyright, patent, and other intellectual property laws. But it is not always clear who should own data, or whether and to what extent the traditional classifications of property law should apply to it. Should the data that governments collect, for example, be owned by those governments, or should it be owned by the general public? Should doctors and teachers and retail businesses and governments own the data they collect about their patients, students, customers, and citizens because they have taken the time, and expense, and trouble of collecting it, or should their patients, students, customers, and citizens own it because it is – after all – data about them? Should they all, perhaps, share in its property rights? What about the information that someone infers or derives from data that was collected by someone else? When and under what conditions should data become "IP-able"? And what, in any event, should this mean about whether and to what extent others can use it?

Here, the participants said that data is power; that it is "the new oil", and that there are typically two values, or principles, associated with sharing it. The first is that the data owner, regardless of who we decide it might be, should be able to decide how it is used, including who to share it with, how to share it with them, and under what conditions. The second is more about the need to apply a joint ownership model to balance the shared rights that governments, businesses, and citizens have to use data. Data owners, according to the first principle, may decide to share their data freely with certain people and institutions if they like them or approve of what they want to do with it, so that people may be free to access data so long as they conform to what its owners think, want, and believe. Alternatively, they may decide to sell it – or the right to use it – for their own personal or institutional benefit. But the idea that data and information is something that can be owned is a conceptual construct of our own historical epoch that is as different and novel as talk about the ownership of the planet. And it is not at all clear whether data should be regarded as intellectual property that individuals and institutions can, or should, be allowed to own. A lot of data is currently withheld from the public, and some institutions are especially secretive in withholding it, because the people who control and process it expect eventually to benefit from it. Governments may be the largest users of data, and perhaps the most secretive. But there are also commercial users, such as corporations, who have obvious financial interests in keeping it secret. The value of data obviously depends upon its use, including the use in keeping it secret. Our participants, insofar as this is concerned, said that most of the benefit of data should ultimately end up in the hands of consumers – both individual consumers and business consumers – and that it should generally have a value in empowering them. But they also said that data doesn't know how it will or might be used, or for what purposes, and that the same data may obviously be used for many different purposes. They said that the amount of data that governments and industry have collected about individuals, and the inability of those individuals to know that it exists or control how it is used, has created an unhealthy asymmetry of power. They thus voiced concerns about establishing monopolies on information and charging excessive rents for using it. They said that it is very difficult to apply a sole intellectual property ownership model when it comes to data, and that the discussion should be more about shared rights – who has them and whether it possible to come to agreement about them – than sole ownerships.

5. The Transparency of Data

The transparency of data was also a governance concern for our discussion participants. This issue is of public concern because it may now be possible to share all of the data that we collect with anyone that wants to use it. But the question, of course, is whether we should. And if not, how much of it should we share, with whom, and under what conditions? Here, some participants said that we should all have equal access to data, and that governments and private entities have a moral, social, and economic obligation to share whatever they know with whoever wants to know it. They seemed to regard equal access to data as a fundamental right. They said that making data inaccessible or obscuring its meaning is both immoral and one of the best ways to control people and to keep them powerless, and that it is also socially and economically counter-productive. They thus argued for a policy they called "radical transparency". They said that we cannot know in advance what the next ground breaking innovation will be, where it will come from, or who will make it. That restricting the availability of data to certain users inevitably pre-selects who can profit from it, that it does so regardless of whether the profit is economic, political, or social in nature and that everyone should thus be able to access any piece of information they want so long as we can collect it. They said that whenever data is kept secret, its owners have a monopoly on evaluating it and drawing conclusions from it, and can thus exercise tight control over its public message. That there are both individual and corporate interests in transparency and that radical transparency could also help to improve the quality of data; for if the data is available to everyone, then everyone can analyze it and find that different conclusions can be drawn from it. They also felt that this would go a long way toward leveling the political, economic, and social playing field by shifting the power equation so that everyone has the opportunity to profit from the data revolution, though the powers that be do not want to hear this.

Others were not so sure. They generally agreed that data should be more transparent, and they generally thought that this will be the way of the future, but they worried that radical transparency might compromise our privacy, our safety, and even our integrity. They thus worried about its implications for confidentiality provisions in contracts, trade-secret law, non-disclosure agreements, the privacy of court records in child abuse and other sensitive cases, and the "right to be forgotten" in Europe; all of which seem to pose exceptions to any moral, social or economic obligation to share what we know with anyone who wants to know it. They said that radical transparency would also give our enemies access to information essential for national security. They pointed out that we live in a competitive world, that individuals and corporations have invested time, money, and effort to create the information technologies which have made big data and its collection possible, and that they should have a right to profit from it more than others. Others, however, insisted that corporations should share their profits from the information they collect about individuals with those individuals, and that there needs to be a balance between what data can be made available, and what data requires an individual's consent prior to being made available. Still others said that radical transparency would not alter the power equation at all, but only shift power to those who have the mathematical skills and technical wherewithal to deal with big data. But they all wondered about how we would make all of the data that has been collected transparent.

Most of the discussants thought that radical transparency was a non-starter as a policy proposal, but they all thought it was a good place to begin a discussion about transparency, since it would inevitably lead us to think about what kinds of data should *not* be made transparent, and about who we should share information with and under what circumstances. And this led to an interesting discussion about "meaningful transparency", which introduced the dimension of the data user's interests and the idea that data should be presented to users in a way that is easily accessible, understandable, and consumable.

6. The Privacy of Data

The privacy of data was another major governance concern for our discussion participants, though it is not at all clear whether they thought there is a meaningful distinction between the privacy of data itself and the privacy and privacy rights of the people to whom the data refers. Some participants said that data security can affect the privacy and privacy rights of individuals, but that data itself does not have privacy. Others said that people can decide to keep data secret, or try to retain property rights for it, and that there is no reason not to talk about the privacy of data and private data if and when they do. Some of our participants' concerns about privacy were the flip side of their concerns about transparency – especially when it came to the data collected about individuals involuntarily – since they thought that the more data that is collected about individuals, and the more transparent it is, the less privacy those individuals will be able to retain. This, however, was just one view. Others said that transparency may actually enable individuals to enjoy more and better privacy, since if people know what data exists about them, then they can exercise more and better control over it, and that this is the reason why it is one of the core principles of the Consumer Privacy Bill of Rights. And still others said that privacy is a thing of the past and that we need to "get over it" in order to reach our full potential.

Some participants raised questions about whether and to what extent information that people give away about themselves passively is voluntarily or involuntarily disclosed. They said that it is easy to get lost in a quagmire of concerns about privacy, that such concerns are generally personal, moral, and legal issues that take us away from questions pertaining to the acquisition and use of data for social good and that we can easily lose sight of the benefits we can gain through sharing data and information if we do. Others disagreed; they said that many people are very concerned about people and institutions and companies and governments having too much information about their everyday lives, especially if they lack "digital self-determination" and are not able to control how they might appear to those who might use it. They also expressed concerns about how the data about them will be used. And while this may, once again, be a concern about the use of data, it is specific to possible uses that violate their privacy. They thus worried about identity theft. about the growing sophistication of cyber attackers, and about discriminatory practices such as the use of private medical records to increase insurance premiums, or the possibility of losing their jobs due to unwanted disclosures about their personal lives. They also worried about disruptions that such disclosures might have upon their personal lives and relationships. They said that there is always a concern that the data collected about people is "a ticking time bomb that may go off at any moment", revealing things they would prefer not to be known.

Such concerns raise doubts about the confidentiality of personal information. For when all is said and done, it is impossible to prove that well intentioned attempts to ensure an individual's privacy by anonymizing his or her personal dataset cannot be later reversed by combining it with other datasets, including datasets that do not exist at present, or might not be transparent at present, but which might become both available and transparent in the future.

7. The Moral Nature of Data

Our participants also raised a number of moral concerns about the collection and use of data aside from those pertaining to privacy. Moral questions are generally about right and wrong conduct, and about how we should or should not behave. Our participants said that people sometimes think of data itself as being good or bad: good, presumably, if it is used in ways that they perceive to be good and bad, presumably, if it is used in ways that they perceive to be bad. But they also said that we collect data to gather information, that simply gathering information is morally neutral, and that there is nothing inherently good or bad about doing it. The concerns pertaining to privacy discussed earlier are moral concerns about whether it is right or wrong, good or bad, to collect and publicize certain kinds of information about people, especially if doing so might cause them harm. But there are also questions that arise from the possession of information where that information may be used to *prevent* harm to people. Do we, for example, have a moral obligation to try to prevent crimes from occurring if we are able to predict with a high degree of probability where and when they are most likely to occur? Do we have a moral obligation to outlaw certain foods if we are able to predict with a high degree of probability that people will suffer disease as a result of eating them? What if we can predict with a high degree of probability that people with certain genetic makeups will contact serious diseases? And what if we can predict with a high degree of probability that students with certain backgrounds will not do well in school, or are more likely to commit crimes, or that they are more likely to place an economic burden on the state?

Questions like these can be multiplied at length. And the answers to them are not always clear. Some of our discussion participants thought that having convincing data which would allow us to make highly probable predictions about such issues places a moral obligation upon us to do something about them. But others disagreed, and still others said that we have other moral obligations that take precedence, such as the obligation to uphold the freedom of individuals, equal opportunity, and equality, and that living up to them might even mean that we should stop collecting certain kinds of data which might lead to their erosion.

Another, and somewhat different, moral concern that our participants raised about data, and especially big data, was whether and to what extent the very use of numbers and statistics to describe the human condition and the situations of different human beings might somehow dehumanize them in our eyes, so that abstract numerical talk about the various categories into which certain people may fall might encourage us to regard them as somehow less than human. Does talk about the 1%, or the 99%, or the 0.001%, or the 47%, in and of itself make us feel less empathy and moral concern for the flesh and blood people who fall under these categories? Does it lead us to feel less empathy for some people, and more empathy for others? Does it affect the ways in which we think we should think about them, or the ways in which we think we should act toward them, or what we think we can and should reasonably expect from them? And if so, then should we regard the reduction of flesh and blood human beings to the abstractions which talking about numbers and statistics may sometimes suggest as somehow immoral in itself?

8. The Public and Private Good of Data

Concerns about the moral nature of data are closely related to, but somewhat different from, concerns about the relationships between data on the one hand and the public and private good on the other. Our discussion participants, in any event, voiced concerns about these relationships as well. They thus asked how we can leverage the possibilities of our new technologies and data for the public good; how international organizations and governments can fund the collection and use of data for maximum public effect and how public and private organizations can remain agile enough to make the best productive use of their growing stockpiles of data and the rapid but unpredictable evolution of our technologies for collecting it. But they also noted that the public and private good are often in conflict with each other, and raised questions about how to determine what is and is not a public or private good – and indeed about how to determine the boundaries between the public and private realms in an era of big data. Is an individual's illness a public or private matter? What about his or her sexual behavior? Should people be required to disclose their medical and sexual histories in order to receive publicly financed medical treatment? Should they be required to disclose their medical histories to help health workers and researchers identify new disease trends and outbreaks? What about a person's criminal record, or a person's educational record? And who, in any event, would or should determine what is and is not a public good?

Our participants noted that people in China and India are much more willing to accept invasions of their privacy if they are convinced that it will help the public good than people from Western countries, and that people in Great Britain are more willing to accept cameras in public spaces for the public good as well. But they also noted that going too far down this path could all too easily transform a free and open society into a police state. They thus worried about the wisdom of collecting data about everyone in the name of the public good when it can so easily threaten another public good. They said that we should develop a code of ethics similar to the Hippocratic oath (which provides guidance to doctors treating patients and cautions them to "do no harm') to govern the public and private use of data; a code which would specify criteria to evaluate and determine fair and appropriate uses of data for the public and private good. But they also worried about who would actually determine what is and is not a fair and appropriate use, and they said that it is interesting, at the very least, that the public expects private companies to act in the public good, since the very essence of a private company is that it can do what it wants and not be subject to the norms of the public domain, while at the same time wanting governments to assume roles that used to belong to the private domain.

This last point is very important. If governments lack the geographic reach to ensure that data will be used for the public good, and corporations lack sufficient returns on investment to do it, then we will need to discuss what kinds of institutional constructs it will take to do it.

9. The Effect of Data upon Data

Our discussion participants also voiced concerns about the effect that data has upon data itself. They said that viewing data as static can be problematic, since knowledge of the data may well change the ways in which people act, thereby rendering the data that we have incorrect or useless for predictions, and that we need to find some way to account for this. They said that data might be subject to something like Heisenberg's Uncertainty Principle in that we may have an effect upon something simply by trying to measure it, or George Soros' reflexivity principle in that what people may do in light of their knowledge of a prediction may have both positive and negative effects upon its outcome. Statistical data indicating, for example, that on average, Americans owe $15,000 in credit card debt, or $154,000 in mortgage debt, or that the average student loan debt is $34,000 may lead some people to borrow less and others to borrow more, and further predictions that these numbers may increase or decrease in the future may have similar effects. They also said that we need data about the successes and failures of collecting data, and about the successes and failures of interventions that are based on the data we have collected. And they said that the reliability of data is just another piece of data, that it is generally in the eye of the beholder, and that which data is or is not reliable is a question not only about how it is collected but also about the perspectives that people have on it and the purposes for which they want to use it – so the idea that the absolute measures recorded are reliable in and of themselves is usually a product of what you want to do with them. They said that we can always raise scientific questions about the reliability of the ways in which we have acquired data, but that we often rely on data in order to accomplish something, whether it is theoretical or involves practical action in the world, and that we need to know that the data has been collected in the same ways in order to believe in its reliability enough to act upon it. But they also said that reliability of data and its consistency are two different things. They said that there is always institutional competition among the different agencies that collect data, that collecting data in different ways or from different sources may easily affect its consistency, that this is also data about data, and that changes or inconsistencies in the way we calculate GDP, for example, may affect everything else. And they said that the quest for consistency can itself be problematic since the desire to do something with it may tempt people to inject errors into the data in order present a more harmonious picture by "correcting" for its inconsistencies.

Entr'acte

These are the nine governance concerns that our discussion participants raised about the nature and use of data. I think that they clearly overlap in many ways and that each of them must be addressed in one way or another before citizens will feel comfortable about the proposed data revolution and the agenda for "The Post-2015 Sustainable Development Goals". Exactly how we should address them is another question, and one that will require careful thought and experiment. But in the remainder of this chapter, I would like to focus my attention on a specific use of data that concerns me; namely, the use of data to justify or promote public policy proposals. And in what follows, I will argue that data should not be used in this way.

10. Public Policy and the Critical Use of Data

Earlier, I noted that our discussion participants were concerned about the use of data to promote public policy possibilities and justify policy decisions that were actually based on different grounds. Here, I would like to go a step further and question whether we should try to use data to justify or promote policy decisions at all. Many people today talk about the need for "evidence-based" policy decisions. I agree with them, but I think that the idea they have about evidence-based decisions is different from my own. The difference I see does not lie in the *use* of data and its resultant analytics for making policy decisions, but in the *way* they are used. Proponents of evidence-based public policy decisions think that making public policy can and should involve data and empirical evidence that has been rigorously collected and objectively established, preferably through the rigorous use of scientific methods. So far, so good, but many of them seem to think that data and empirical evidence can and should actually *determine* policy; that the facts are, as it were, all that should matter when it comes to making policy decisions, so that once we have them everything else should follow. This idea of evidence-based policy decisions is very intuitive, and the image is very clear. They want to build our public policy on a firm foundation instead of sand – in the same way we want to build our skyscrapers on a firm foundation instead of sand – and they think that data and empirical evidence will provide the firm foundation. They would, in this way, like to replace ideologically driven public policy decisions with data driven public policy decisions. But I think that this idea and the beliefs that motivate it are fundamentally flawed.

There may, in the nineteenth and early twentieth centuries, have been reason to hope that empirical evidence could one day underwrite the certainty of scientific theories in the social sciences, just as it had underwritten the certainty of scientific theories in the natural sciences. In those days, philosophers and scientists generally regarded scientific knowledge as justified true belief. They thought that it was objectively certain, and they dreamt of the day when scientific knowledge would help to both predict the future and resolve our social problems in the same way that theories in the natural sciences have helped us to predict the future and explain the natural world around us. But a hundred and more years of closer philosophical scrutiny suggest that data and empirical evidence simply cannot do what we once thought they could, even in the natural sciences. Philosophers of science as diverse as Karl Popper, Thomas Kuhn, and even, eventually, Rudolf Carnap all recognized that scientific theories cannot be underwritten or justified, by "the objective facts"; that no universal theory can be justified by empirical observations, no matter how good or how many there may be, and that even the facts that would underwrite them are inherently both theory and value laden. Most philosophers and scientists today have thus come to accept that scientific theories are always underdetermined by the evidence; that no finite amount of empirical evidence can show that a scientific theory is true, or even probably true, no matter how much data there is and how properly it has been collected, and that scientific theories are thus inherently and irremediably fallible and always subject to revision – which is, perhaps, the reason why we hear so much today about "the consensus of belief within the scientific community" instead of the evidence upon which one might hope it is based.

But it's not just the fact that scientific theories are by their very nature irremediably fallible and always subject to revision. Public policy is simply not based solely or even primarily upon empirical facts. I know that proponents of evidence-based policy think that this is one of its primary problems, but public policy decisions are rightly based to a much greater extent than most proponents of evidence-based policy generally realize, or would like to admit, upon the non-scientific concerns, beliefs, values, goals, interests, and priorities of individuals and institutions, and these are logically independent of empirical facts. We may thus often, and rightly, decide to pursue a policy course even when – and perhaps especially when – its opponents think that the scientific facts suggest a different one. We may call these beliefs, values, goals, interests, and priorities a philosophy, as opposed to a science, especially if they are rigorously consistent and well articulated, or we may call them an ideology if we strongly disagree with them. Scientific evidence may and should inform them in the policy-making process, but it cannot replace them, no matter how good it may be. And the reason why it cannot replace them is also easy to see, for the very first task for practitioners of evidence-based policy making is to clearly define an over-arching policy objective, such as "The Post-2015 Sustainable Development Goals", which they want to achieve. Such objectives can be very broad or very narrow. If they are too broad, their success in garnering general support may not translate into similar support for the more specific policy proposals necessary to implement them. And if they are too narrow, they may not garner general support at all. But here, the point to be made is that we typically define such over-arching policy objectives in an effort to address our social problems. But what we perceive to be a social problem typically reflects our own philosophy – or our ideology if we strongly disagree with it – and not simply objective information, empirical evidence, or data.

Data may suggest that something is a fact. And that may be something, even if we believe that all facts are both theory and value laden. But it cannot tell us whether or not it is a problem – let alone an important problem – to solve, or give us a policy for addressing it. The very same fact,say, for example, that the average income for a family of four is approximately $50,000 in the United States – which I found on *Wikipedia* and assume to be true (more or less), even though I have found many other "facts" on *Wikipedia* which I know to be false – may be used to support diametrically opposed policies, depending on the different governance concerns, beliefs, values, goals, interests, and priorities that we may have. We may, if we think that $50,000 is a sufficient income for a family of four, decide that there is no problem and nothing more needs to be done. Or we may, if we think it is not, decide to implement programs that will make up for the shortfall.

I have already said that many of our participants felt that data is often collected to justify public policy decisions that have already been made for other reasons, that the evidence used to justify or promote such policy decisions is often cherry-picked, that the very practice of using data to justify or promote such decisions encourages cherry-picking, and that evidence running contrary to the pre-selected, favored policies is often ignored or given short shrift. I share these concerns and think that using data to try to justify or promote policy decisions in this way, no matter how good it may be, can have a corrupting effect that only makes opponents more skeptical of it. So I would like to suggest a different idea of evidence-based policy, one which involves a different and more critical use, and in particular a more self-critical use, of data.

My own idea is that we should not use data and scientific evidence, no matter how good it might be, to try to justify or promote a policy proposal, but only to criticize one. Our governance question should not be whether the evidence justifies the proposal in question, but whether and to what extent it conflicts with it.

It’s not that evidence is irrelevant; flawed as the evidence may be, it is always better to take it into consideration than to ignore it. But that said, the evidence can, at very best, only corroborate some of the concerns and beliefs that motivate a policy proposal – the ones related to empirical facts – by suggesting that those concerns are well founded and those beliefs are true. The proponents of a policy, however, *already* believe that their concerns are well founded and their beliefs are true. So the evidence they put forth is not intended to convince themselves, but to convince opponents of the policy and those who may be still on the fence. And while such evidence may persuade some of those who are still on the fence, and occasionally even some opponents, people who are strongly opposed to the policy will most likely remain skeptical. For we can always have more and better evidence, and we can always invest more time, effort, and money in collecting and evaluating it. But even if we do, the evidence will seldom, if ever, be enough to convince a policy’s opponents, especially because they know that it may be, and often is, cherry-picked. And even if it somehow convinces the opponents of a policy that the concerns of its proponents are well founded and their beliefs are true, they may very well continue to oppose it because it simply conflicts with their own values, goals, interests, and priorities.

These values, goals, interests, and priorities are more philosophical in nature. They cannot be refuted by facts and evidence. They are typically the real decision factors at issue and they vary from person to person. And they will always remain untouched by the data and empirical evidence and be in need of negotiation. I once heard a woman at a global warming conference express the point in a stark and stunning way. She said “Even if the so-called scientific evidence for ‘man-made’ global warming and the predictions you base upon it were entirely true, which I personally do not believe, I would still oppose your carbon policy because it is simply not in my interest.” This is not an isolated example. I do not think it helps to claim that there are some policies that are in everybody’s interest. Nor do I think that it helps to scold or vilify people and institutions that oppose a policy that is not in their own interest any more than it helps to praise people or institutions that support a policy that is in their own interest. It simply underscores the fact that interests and evidence are two different things, and that the evidence is not the only – or even the primary – factor involved in making policy decisions.

This is a point that those who believe in evidence-based policy decisions should take on board.

It is far more effective when we use data, facts, and empirical evidence to criticize a policy rather than to justify or promote it. Confirmations of a theory are a dime a dozen, but failing to find evidence against a belief after making serious and sincere efforts to find it is typically more convincing to those who may be on the fence and might be swayed by it.

Finally, I think there is, and will always be, rich and fertile ground for serious public discussion about the competing philosophical values, interests, goals, and priorities that motivate our governance concerns. These values, interests, goals, and priorities will always play a much greater role in policy decisions than scientific facts and evidence, and policy makers should till that soil as best they can instead of dismissing them out of hand as ideology.

Coda

I want to end by emphasizing that the ten governance concerns discussed in this chapter are essentially philosophical issues, that they will probably always be with us, and that we will probably need to consider them over and over again as our datasets become larger and more varied and as our computing facilities become faster and more sophisticated. I also want to emphasize one of the considerable dangers of the big data revolution: namely, that we may begin to expect too much from the large datasets we collect and the new information processing technologies we create, and too little from ourselves. The halcyon days of scientism that we call the nineteenth and twentieth centuries are now at an end, not because our science today is any worse, but because we have, as a society, become much more realistic about what science can and cannot do, even if some people still think that it can prove that our theories are true, or that "the science is settled", or "the debate is over". Some people, no doubt, will always deify science as "The Truth", and vilify those who have the temerity to question it as immoral. And scientism itself may even have a resurgence in the era of big data, even though some people think it spells the end of theory. But using data to test scientific theories and justify policy initiatives is a very different thing from using information technologies to collect data and find patterns of association or correlation in it. And the fundamental epistemological realities will always remain the same.

Much of our success in using data will always depend upon the questions that we ask at the planning and design stage of a project. And we should not expect to produce helpful results at the data processing stage if we ask the wrong questions at the start. But regardless of whether we ask the right or wrong questions at the start, there will always be tendencies at work which will add to the difficulties of using data. One of these is the tendency to cherry-pick data to find the patterns and correlations that we want to find. Another is the tendency to interpret data in ways that confirm our preconceived beliefs and to ignore data that conflict with them. And a third is to polarize policy debates by suggesting that those who are unimpressed by the data that impresses us must be either stupid, malicious or immoral. When these tendencies work together, as they often do, they inevitably create a climate that is unhealthy for the proper analysis, reporting, and use of data, and one which is, indeed, destructive of healthy public policy discussion in a free and open society.