

Harvard Data Science Review

Rejoinder: The Present and Future of Data Science in Society

Sabina Leonelli

Published on: Jan 29, 2021

DOI: 10.1162/99608f92.fc216595

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

I am delighted for the opportunity offered by HDSR to have a wide-ranging conversation over the role of data science in the COVID-19 pandemic, and I am grateful to all the discussants for their generous comments and insightful expansions on my original piece. In what follows, I identify and discuss four key topics that emerged from this conversation, with the aim to prompt further dialogue on the use of data science to tackle the pandemic as well as on the future of data science and its role in society beyond the current crisis.

From Clinical Spaces to Planetary Health

I will start from [Ogburn's excellent discussion](#) of clinical decision-making as a 6th imaginary for data use: one which has received little attention and thus largely failed to inspire health-related data practices, at least so far. As Ogburn notes, this 6th imaginary needs to include concrete means to coordinate randomised clinical trials (RCTs)—and related standards, design and data infrastructures—around the world. In the absence of such coordination, a large proportion of the trials set up to test possible COVID treatments produced 'inactionable' data; and drugs have been administered to COVID patients in the face of little or mixed evidence of their effectiveness. Ogburn provides important advice on how to avoid such dangerous failures, such as investing in international data infrastructures (a point I shall come back to throughout this text), applying stricter methods to decide which clinical trials should be going ahead, and rewarding researchers for contributing to large multi-sited trials over leading local trials of relatively little power and significance. At the same time, she evokes multiple audiences who should be consulted in the design and regulation of such a data sharing system. In her words: "we need data scientists to work with bioethicists, regulators, epistemologists, clinicians, and researchers to develop a framework for balancing, weighting, quantifying, and aggregating evidence across different domains such as animal models, biological and *in vitro* experiments, observational analyses, and exploratory RCTs." While she does not mention this explicitly, I think this involves two important conceptual moves for the production and evaluation of medical evidence as a whole.

First is the need to expand existing understandings of clinically actionable evidence beyond RCTs. RCTs have long been marked as a 'gold standard' for medical evidence by the evidence-based medicine (EBM) movement, the pinnacle of the 'evidence pyramid' typically employed to depict the various forms of evidence available to (see Figure 1). The pandemic has given reason to re-iterate the importance of well-run RCTs, and yet also it raised an additional question: does the recognition that RCTs are effective in producing reliable evidence necessarily involve mistrust of other data sources? I would argue that this is not the case, especially in a pandemic situation where observational data coming from the medical frontline has provided information of great and immediate significance for

clinical decisions. Ogborn is of course right to note that much of such evidence is inconclusive and problematic when considered in and by itself, and RCTs retain a crucial role in verifying the effectiveness of proposed treatments. At the same time, observational data and experiences from medical staff play a key role in fostering the discovery of potential treatments. They are at the bottom of the evidence pyramid not because they are “bad evidence”, but because they are foundational to all other forms of data collection. In this sense, biomedical researchers confronted with an evolving emergency would benefit enormously from an effective system to collect and compare a broader range of data sources collected on the medical frontline as well as the lab. Publishing such data in medical journals is a relatively slow and ineffective way to facilitate evidence integration, and precludes making datasets searchable and interoperable. Hence the need to develop and implement a comprehensive strategy to identify, share and analyse data of potential relevance to public health. This should constitute the top priority for funders and policy-makers looking at strengthening preparedness for future disasters (see also the ongoing work and recommendations of the Research Data Alliance COVID-19 Working Group: <https://www.rd-alliance.org/groups/rda-covid19>).

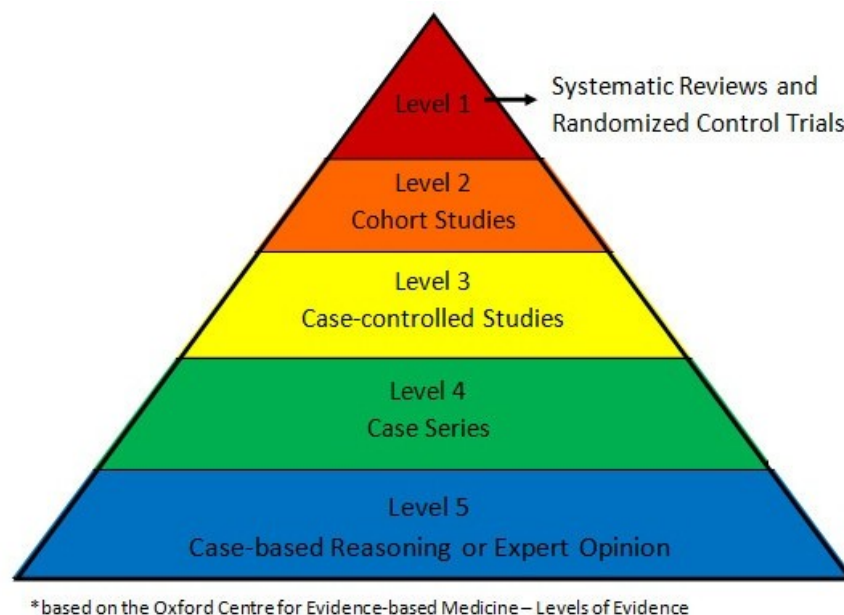


Figure 1. The pyramid structure of sources of evidence according to evidence-based medicine, where the pinnacle of the pyramid comprises the most reliable source. Whether or not one agrees with this ranking, the question to ask is whether and how different levels of evidence depend on each other (from <https://sites.google.com/site/evidencebaseddd/types-of-pyramid>, CC-BY).

The second conceptual move is the relation between the technical discussion around which data should inform clinical decision-making and how, and broader questions around what is taken to constitute the ‘clinical space’ and why. Ongoing efforts to tackle the pandemic are contributing to the recognition that the clinic is not a self-contained space for administering treatment to patients. Clinics are

interconnected with the rest of society in myriad ways which are extremely difficult to contain and control. Countless forms of data are produced and haphazardly disseminated. Opinions are formed and circulated among staff and patients alike, sometimes making their way into social media and news outlets. Various forms of waste—including infected or radioactive materials—circulate in and out of hospitals. Human traffic is also considerable. People enter the clinic for widely different reasons, ranging from in-patient visits to regular health checks; and clinical workers (including cleaners, technicians, nursing staff, doctors, and administrators) have regular contact with their own social networks. As famously argued by Michel Foucault (1963), among others, attempts to curb and control these flows of materials, knowledge and agency have shaped the clinical space since its 18th century emergence.

In the case of COVID-19, the clinic's unavoidable interrelation with the world has been curbed through efforts to minimize exposure to the clinical space, as exemplified by full-body PPE and extreme social distancing. This has proved extremely demanding on all the humans involved. Stories of doctors living apart from their children and patients being forcibly separated from loved ones (including those receiving care for non-COVID diseases or for conditions that have nothing to do with illness, such as pregnancy) have highlighted the sense of alienation underpinning the social distancing enforced by the pandemic upon large sections of human society, as well as the crucial role of the clinic as a crucible of social experiences. Images of clinicians holding up a phone for critically ill patients to speak to parents, children or spouses, so frequently seen on the news, have now entered the lived experience of many of us as cases multiply particularly in Europe and North America. Data scientists working on pandemic responses need to acknowledge and investigate the sociality of the clinical space and what this means for the selection and analysis of health data. Like other spaces of social containment, such as prisons and immigration camps, clinics exemplify the difficulties of isolating the sick and the devastating emotional, physical and social effects of COVID-19. Perhaps more overtly than other spaces of social containment, clinics also offer hope through treatment, thus producing countless opportunities for data scientists to constructively engage with the vast information being produced in order to improve the effectiveness as well as the humanity of clinical care.

Public health in this broad sense affects all forms of sociality and related spaces. Reflection on the porous boundaries of the clinical space, and the roles that data science could play within and beyond them, is therefore compatible with recognizing the role that data science could play in other settings, ranging from prisons to schools and workplaces. Hence, I fully agree with [Kolaczyk et al.](#) that “while the contributions of data science can certainly be scoped around the healthcare and public health needs, it should not be limited to them.”

This point beautifully complements the analysis of the clinical actionability provided by Ogburn, and I certainly never intended to imply otherwise: my proposed imaginaries 4 and 5 (using data to address

logistical and social need) involve data-intensive research spanning virtually all social domains and contexts, from travel to urban design, employment rights to racial inequity.

Kolaczyk et al. rightly emphasize that the insistence on a holistic, transdisciplinary, and socially-engaged approach to disasters, including pandemic, is not new. There is a long history of researchers, activists, and international organizations repeatedly asking for substantive and wide-ranging investment in preparedness over the last century. What is perhaps more novel, at least in terms of public perception of risk, is the increasingly strong overlap between health concerns around the pandemic and environmental concerns around humans mistreating the planet and its non-human inhabitants. As the clinical space expands into society at large, the pandemic has made ever more apparent how public health depends on the wide variety of creatures that populate our shared habitats. Truly, the idea of *planetary health* is taking hold: we are all in this together, not just as human race, but as living organisms co-existing in a complex ecosystem whose resilience is being tested to its limits. Maintaining this awareness while working towards improved data infrastructures and evidential strategies for clinical decision-making is a crucial requirement for 21st-century data science.

From Imagination to Socially-Robust Action

[Louissaint's piece is a call to act now](#) since, as she puts it, “the middle of a pandemic response is not the time for future gazing.” Her powerful contribution identifies several critical concerns that data science can immediately help to address and mitigate, including the spread of misinformation, housing and economic insecurity, and the unequitable impact of COVID-19 on minorities and vulnerable parts of the population among many others—thus stressing the importance of addressing social needs, which I had identified as the fifth imaginary for data use. I certainly agree with her that time is of the essence in grappling with issues that are causing untold amounts of suffering every day that goes by. As I write, COVID-19 infections—including those caused by newly evolved variants—are inflicting record-breaking death tolls and long-term disability levels across the American continents, Africa, and Europe. When calling for data scientists to consider the imaginaries underpinning their work, I am therefore not asking researchers to waste precious time, but rather *to secure means to assess and question the social robustness of the solutions that they develop*.

Time and time again in the history of science and technology, tools developed with the best intentions and under enormous pressure to address urgent social challenges have turned out to have questionable long-term effects (think only of nuclear power, genetic engineering and facial recognition systems as glaring examples). This cannot be avoided: nobody can control precisely how a given application or technique will be used once it starts travelling across contexts, and it remains hard to determine when the risks of a given technology may outweigh its advantages. However, the rush to offer solutions cannot be used as an excuse to avoid any consideration of the social embedding of those solutions. In other words, the scientific imagination needs to include the social, and to do that beyond

mere wishful thinking about ‘best scenarios’ in highly idealised settings. Many researchers are committed to devoting their efforts to the public good; fewer researchers attempt to explicitly articulate and question their own assumptions about what the public good involves, for whom, and under which conditions. It is this latter effort which, I argued in my paper, requires conversations beyond one’s own discipline and habitual peers. Just like the scientific imagination is grounded on empirical insights about the way the world is, so is the social imagination for how scientific outputs may affect the world. For any given context, there will be a cluster of experts (which may include social scientists as well as humanists, community representatives, professional figures and civil servants, among others) who possess an evidence-based understanding of that specific context. This knowledge can and should be mined when assessing how a given technical solution may or may not work within a given socio-cultural setting. Furthermore, history demonstrates how creative uses of data can be spurred at least as much by confronting the concrete characteristics of real-world situations as they are by pursuing highly idealized visions of the future of humanity. Consider the frequency with which revolutionary discoveries have emerged from highly applied data analysis, ranging from the rise of computing from WWII cryptography to the birth of epidemiology from the mapping of cholera transmission on the street maps of 19th century London.

The attention paid to tracing apps in the first few months of the pandemic is a case in point. This was a situation where the opportunity to combine smartphone use and data science to track the spread of contagion—and alert individuals who may be at risk—was widely perceived as key to pandemic containment. And indeed, this solution seems to have yielded public health benefits when it was integrated from the get-go with other public health measures, including traditional contact tracing via local health authorities and individualized medical support (as was arguably the case in Ireland and Germany; see Jee, 2020). In other words, this solution proved useful when it was immediately and sustainably complemented by meaningful, well-informed action on the ground. In many cases, however, this did not happen. Technology was presented as working in and of itself, and little effort was devoted to aligning it with social and medical services, and/or to consult on its broader impact on different social sectors and communities. Governments—including the UK, US, and Italy—invited their citizen to download tracing apps without securing support from local health authorities and consulting potential users, and, indeed, without integrating those tools into a broader public health plan of assistance and support for individuals at risk of contagion. The main ethical concern raised around the effectiveness of contact tracing apps was privacy, with Apple and Google coming to the rescue by proposing an additional technical fix—an anonymous, Bluetooth-based alert system geared to make it impossible to use such apps as surveillance tools. Yet, again, this way of solving the privacy issue was not scrutinized for its broader consequences: the app had to be launched as quickly as possible to ensure effective tracking and tracing of the disease, and an empirically-informed

evaluation of its social implementation was seen as wasted time in the midst of an unfolding emergency.

As it turned out, the rollout of the tracing system proposed by Apple and Google actually made it harder for health authorities to intervene effectively. Its privacy-preserving features got in the way of efforts to verify who had been alerted by the app and provide adequate support to those individuals, their families, and employers; helping individuals to distinguish false alarms from genuine risk; and using such information to strengthen understanding of contagion locally and nationally, thus informing further interventions. This in turn generated a widespread loss of trust with the app among the citizens who were supposed to rely on it. Hence not only have many tracing apps—and related investments—not lived up to their hype as key tools for the pandemic response in Europe and North America, but their problematic implementation has fostered disillusionment with governmental interventions and scientific advice, at a moment where misinformation and mistrust of authorities were already having crippling effects on efforts to mitigate the spread of COVID-19. Public health experts, social scientists, and various citizen groups had warned against this outcome, and indeed, initiatives such as StopCovid in France attempted to avoid it by setting up a home-grown system intended to safeguard citizen rights while also enabling support by local health authorities (Krige and Leonelli, 2021). It is early to evaluate the overall effectiveness of these initiatives, and whether and how contact tracing through smartphones actually contributed to the pandemic response, but the rush to implement such systems as immediate technical fixes regardless of the social context does not seem to have paid off.

The scientific imagination around tracing infections failed to adequately consider the social, and, in doing so, damaged public trust in both government and science. This is the kind of rush that we want to avoid, in my view.

From Politics of Data Science to Politics *in* Data Science

This brings me to the role of politics in data science, and the thorny issue of whether and how to split accountability for the success and failures of proposed solutions between, on the one hand, the scientists who decide whether and how to develop them and, on the other hand, the policy-makers who decide whether and how to implement those solutions. [El-Sayed and Prainsack eloquently illustrate](#) the extent to which scientists bear responsibility for their interventions and proposals, for instance when determining what data should be collected in the first place – and thus which aspects of human experience are worth considering for further analysis. As demonstrated by [my own work on data use](#), I agree with their arguments and did not mean to belittle the political role and responsibilities of data scientists (and other experts) in the context of the pandemic and beyond. Rather, I meant to stress the substantive responsibilities borne by policy-makers, which range from their role in shaping the conditions for scientific work, including setting goals and allocating financial

resources, to their translation of scientific findings into social interventions, which requires interpreting those findings in light of specific socioeconomic priorities. Hence my argument that policy-makers should not defer responsibility for their actions to the scientists—a strategy that has proved disturbingly popular among political leaders confronted with the stark consequences of social distancing measures, with many claiming to be “led by science” when asked to justify their decisions (Dupré, 2020).

How to make sense of the complex relation between scientific and political responsibilities and interventions in the context of a major emergency such as the pandemic? I am convinced that political commitments and related accountabilities are present in both research and policy-making, but that they take different forms within these two realms.

While data scientists need to recognize and explicitly address the role played by politics within their research, they have limited control on the material and financial conditions under which their work is supported and on how their work may be used as evidence for policy measures. What scientists do have, particularly those in established positions, is the opportunity to consider how political circumstances and commitments may impact the direction, methods, and outcomes of their research; and to challenge political decisions, especially when they perceive some political readings of their work as clashing with their findings.

This is not quite the same as ‘speaking truth to power.’ For a start, science does not produce absolute, unassailable truths. Rather, science produces evidence and arguments for why certain claims are highly plausible, in the hope to foster an ongoing dialogue over the significance of those claims, including the ways in which they could be challenged. In this sense, the prediction that a given vaccine, when administered in specific doses, is effective in over 90% of cases is not a statement of truth. Rather, it is an assertion grounded in robust evidence and reasoning, and which therefore is more reliable than assertions that have no such grounding. The opportunity to explain and reproduce scientific reasoning and methods, the lack of dogmatism and the openness to critical scrutiny are key ingredients for the success of science as a system to produce reliable knowledge—arguably the best ever devised in order to understand the world. This system also includes mechanisms to identify the value-judgements made by researchers in their work, and make them as explicit as possible so that they can be evaluated as part of the scientific output. This is what many students of science—and some political figures, as demonstrated by [Lord Clements-Jones in his comments](#)—mean when pointing out that science is value-laden through and through (e.g. Douglas, 2009). Most of the researchers working on COVID-19 vaccines are motivated by a strong desire to curb transmission: this is an ethical and political commitment that is openly acknowledged and accepted as an incentive rather than an obstacle to the production of reliable knowledge.

There are, of course, values and commitments that do stand in the way of good science. Recognizing science as value-laden is not a license for scientists to cherry-pick results and ignore ‘inconvenient’ findings on the basis of one’s beliefs and interests. Quite the contrary: it is a recognition that such practices exist, whether by accident or by design, and that the research community as a whole is committed to identifying practices that may be problematic or misleading in the context at hand, and rooting them out. In his discussion piece, [Leslie has beautifully reviewed](#) the development of such awareness throughout the 20th century, and highlighted the importance for scientists to commit to “carrying out the endeavours of science in the public interest” with an eye to social inclusion and planetary health. These views mirror Lord Clements-Jones’ invitation for governments to “lead the way in actively explaining how data is being used”—which as, he rightly points out, is critical to retaining public trust—and for scientists to pay attention to the values instantiated by their work.

What does this mean for data scientists in concrete terms? I will here briefly mention three crucial junctures where data scientists could benefit from an explicit reflection on the values underpinning their technical decisions. A first juncture happens when evaluating whether to participate in a given project or call for funding. It should be clear to scientists that agreement to contribute to a particular research program does imply some degree of acceptance of the overall goals and methods set by the project or funding call in question, including the all-importance choice of admissible data types emphasised by El-Sayed and Prainsack. In turn, the choice of whether or not to contribute to a particular research agenda involves some degree of responsibility over prioritizing that agenda over and above other possible topics. A second juncture happens when designing the research plans and choosing methods and relevant stakeholders. It is often at this stage that decisions are made around which sources and experiences to engage, and whether and how to identify social impact and relevant communities. Ideally, as I argued in my article, these decisions should be made with support from social scientists and representatives of communities with a stake in the potential application of results. With finite resources and timelines, however, hard choices typically need to be made about which interlocutors are best placed to support the project and which sources need to be excluded. These choices are strongly constrained by logistical and institutional conditions, including whether or not certain types of expertise are accessible (e.g., whether representatives of a given patient group are available for consultation) and whether there is scope to engage in such dialogue given the timeframe and resources of the project. Yet the ethos and preferences of investigators also play an important role in decisions around how much and how long to invest in identifying and pursuing such engagements, and how central a role they play in a project. A third juncture happens when researchers consider how to interpret and communicate project results. Like decisions around research design, such assessments can benefit from reflection over responsible and socially sensitive ways to frame and release outputs.

The importance of recognizing the political and value-laden aspects of research, and the ingenious ways in which scientific methods and debates can help to articulate and evaluate them, cannot be underestimated. This is particularly the case since scientific decisions are made in a politically-charged space, where—as so evident in the case of the pandemic—policy-makers have already defined funding priorities and the areas where they wish to see swift progress, often on the basis of predetermined expectations around how research findings may be used to inform social interventions. [As Porter forcefully argues in his comments](#), political action is itself shaped by “an engrained economy of politics,” within which market forces and complex bureaucracies govern the political space to the point of determining the direction and outcomes of much political action. This is also a landscape where, democratic structures notwithstanding, power tends to remain concentrated in the hands of a relatively small elite—an elite that sometimes includes scientists themselves, which is another reason to move away from the rhetoric of science “speaking truth to power.” This enduring reality makes well-intended proposals for the future of technology, such as the “Great Reset” of digital activities proposed by the World Economic Forum to address digital inequity and discrimination (WEF, 2020), sound hollow and unrealistic. Rather than providing an inspiration, such utopic visions of the power of science and technology risk being perceived as a distraction from increasing evidence that the pandemic crisis has augmented digital and social divides all around the globe, with many policies failing to address the multiple forms of “public benefit” required to support widely diverse sections of human society.

In such a context, an assessment of political accountabilities—particularly vis-à-vis scientific findings—remains an indispensable component of any attempt to react to the pandemic without accelerating the self-destructive path of our planet. In the face of self-interested, nationalistic policies sprouting in most parts of the world, it is not trivial to assert that the responsibility for setting out priorities and expectations falls on politicians, whose job involves evaluating the implications of a given intervention for the complex socio-economic landscape at hand—an assessment explicitly informed, within democratic systems, by the core values of their party and the promises made to their electorate. At the same time, scientists can and should challenge policy-makers on the priorities that they set out, particularly in cases where scientific evidence can be interpreted as clashing with policy directions. Anthony Fauci’s staunch resistance to the campaign of misinformation and disruption waged by President Trump, which arguably resulted in hundreds of thousands of American deaths, provided a prominent example of such a challenge. It could of course be argued that Fauci’s role as the director of National Institute of Allergy and Infectious Diseases and a leading member of the White House Coronavirus Task Force placed him in a uniquely influential and socially-accountable position, making this example an exception rather than the rule for scientific engagement with policy. Yet such engagement can also happen through more informal channels, as in the case of the [Independent SAGE advisory committee](#) in the UK. This group comprises prominent public health researchers who felt

dissatisfied with the existing scientific advisory group to the government (known as the SAGE group), chiefly due to the somewhat secretive nature of SAGE deliberations and to suspicions of political interference with SAGE by government representatives. Members of the committee volunteered their expertise towards providing their own assessment of emerging scientific evidence on the pandemic and what kinds of social interventions this assessment may warrant.

The case of Independent SAGE usefully exemplifies the tensions and problems raised by political action undertaken by researchers. On the one hand, this is a clear case of researchers making their values and political stance explicit, by questioning government policies and putting forward concrete alternative proposals (such as closing down schools and extending lockdown measures in the fall of 2020). On the other hand, these researchers have not explicitly acknowledged the political nature of their intervention, preferring instead to highlight the extent to which their pronouncements are grounded on a rigorous and transparent interpretation of comprehensive scientific evidence. Their chosen motto is, accordingly, ‘following the science.’ Given the prominence of a public discourse that confers authority to science on the basis of its putative objectivity rather than its relentless ability to challenge dogmatism with evidence, it is not surprising that such prominent researchers feel that acknowledging their political judgements would undermine the authority and reliability of their scientific work. In my view, this is a symptom of a societal and cultural failure to adequately engage with scientific research and its outputs. As Leslie also highlighted, the tension between scientific objectivity and political advocacy only arises when denying that scientific work emerges from a political context and its interpretation is, accordingly, value-laden. Independent SAGE has stressed that the reliability of its advice can be assessed through its efforts to lay out the underpinning reasoning for external scrutiny. The insistence thus placed on open debate and critical scrutiny reflects an understanding of the scientific endeavour as anti-dogmatic, thus following in steps of prominent 20th century thinkers such as Robert K Merton and Karl Popper. This needs to be complemented by a 21st century sensibility to recognising and acknowledging the political commitments and value judgements unavoidably involved in reasoning from evidence (Cartwright & Hardie, 2012; Leonelli, 2018). Whenever scientists engage in a dialogue with policy-makers, they are explicitly considering some of the social, political and economic implications of their findings, and taking a position on whether such implications are desirable, and why. They are indeed including an imagination of the social into their science: something which is itself, as I argued above, subject to empirical scrutiny.

Looking for Hope: The Crucial Role of Transnational and Transdisciplinary Research Institutions

Can data science offer anything at all to the pandemic response, given the enormous (and enormously underestimated) limits of available data sources and modelling tools, the expanding digital divides that prevent engagement with most human experiences around the world, the entrenched market

logics and auditing systems used to administer essential resources from vaccines to food, and the extensive social barriers separating vulnerable populations from the political, financial and technocratic elite? It may be tempting at such a bleak time to say that the battle is lost, and all data scientists can do is act on the pretense of offering meaningful societal contributions. And yet, as the HDSR special issue on COVID-19 reminds us, data work continues to play a key role in the pandemic response. Effective data sharing and swift collaboration across genomic labs and clinical trials have yielded not just one, but multiple vaccines. Data science underpins key decisions on how to handle the outbreak, providing means to analyze evidence and compare scenarios in situations of real uncertainty where several outcomes are equally possible. Such expertise can serve even the most chaotic political scenarios. Consider again, for instance, the multiple U-turns of the Johnson government in the UK, which has radically changed its own policies on social distancing multiple times (and often from one day to the next) over the last few months.

Such U-turns may well be a signal of political incompetence, but they also exemplify the width of the political space within which data scientists move—including the diverse ways in how data are collected, analyzed and visualized can points to specific measures, and the social dialogue required to assess the broader implications of research results. There is hope in such drama, as well as an imperative for researchers to do all they can to promote socially robust research geared towards planetary health, and to call policy-makers to account when such commitment is betrayed in the name of other agendas.

To help data scientists build on such hope, it is indispensable for governments and industry to investment in institutions, venues and infrastructures that support transdisciplinary research at both the national and transnational levels. This is one aspect on which all participants in this HDSR conversation agree. The rapidity and robustness of data science insights depends on effective collaborations among several stakeholders. This in turn depends on the existence of shared platforms and effective communication networks that are built on the awareness of the social embedding of data science work. As Kolaczyk et al. highlighted, “we require more from data science than relatively isolated contributions and impromptu consortia,” and this means investment in dedicated human resources, long-term institutional agreement and commensurate financial arrangements.

Training for data scientists and their collaborators is a key component of this ambitious project. The various data science training program that have sprouted in institutions around the globe offer further hope in this respect, especially since many of them—including the data science program at my own University of Exeter—include training in data ethics and governance. What is particularly encouraging is seeing capacity for such training increase well beyond the Global North. An example of effective capacity building are the [CODATA-RDA](#) schools of research data science, which take place at regular intervals in different locations around the world, and are especially geared to welcoming students

from low-income countries who are interested in deepening their data science skills. I was temporarily involved in this program in 2015- 2016, when CODATA invited me to one of the early schools (taking place at the Indian Statistical Institute in Bangalore) with the aim to expose data science students to critical data studies and data ethics. Thanks to the many people involved since, including my colleague Louise Bezuidenhout who stewarded teaching on social aspects of data work, this program has blossomed into a well-attended and highly regarded event. Technical and social components of data science are taught together, with serious attention paid to the values and virtues required by different types of data work and the social imaginary that accompanies scientific visions of could and should be achieved (Bezuidenhout & Ratti, 2020).

Another key component of this transdisciplinary approach to data science infrastructures is the development of resources at both the national and the international levels. Kolaczyk et al. demonstrate what this may involve for the United States, where indeed there are resources that can be immediately placed at the service of a strengthened and more socially robust national response. Whether or not this is the case in any one country, it is imperative for every government and national agency to carefully consider how local services may intersect with those set up by other countries, and how data interoperability may be enhanced in ways that are responsible and safe for individuals. Transnational collaboration is fundamental to national goals for several reasons (see also Krige & Leonelli, 2021). It can help to ensure regular exchange of insights and solutions, and thus avoiding an inward-looking, isolationist system that struggles to incorporate and exploit technical developments developed elsewhere. It can foster coordination that maximizes each countries' strengths and helps to counter-balance its weaknesses. And it can help to monitor whether and how data infrastructures are used as means of surveillance, particularly where such use may result in the infringement of human rights. The recently released Digital Services Act (European Commission, 2020), for instance, exemplifies a regulatory mechanism geared to the transnational monitoring of data services and infrastructures. This needs to be complemented by infrastructures and venues that make it possible for such transnational regulation to be implemented.

The road towards data science finding its rightful place alongside other research domains remains long, and it is made ever more arduous by the misleading hype surrounding its capacity to single-handedly boost economies and address global challenges. However, there is perhaps a sense in which data science should be regarded as a 'service domain' after all: not in the sense of mindlessly serving any master or purpose, but rather by exemplifying the interdependence of the research enterprise with all other aspects of the social world. Good data science requires access to both data and relevant metadata, which in turn requires extensive engagement with human experience well beyond academia. Good data science demands venues and infrastructures to care for data, which in turn call for transnational dialogue and support by public and private institutions alike. Good data science involves relentlessly probing the limits and scope of models and algorithms, which in turn demands

exposure to - and critical questioning from - the broadest and most diverse audiences. Last but not least, good data science involves a long-term quest to better understand one's social, political and economic context and impact, and to use that awareness to devise and implement an imaginary of data use.

References

- Bezuidenhout, L., Ratti, E. (2020). What does it mean to embed ethics in data science? An integrative approach based on the microethics and virtues. *AI & Society* <https://doi.org/10.1007/s00146-020-01112-w>
- Cartwright, N. & Hardie, J. (2012). *Evidence Based Policy: A Practical Guide to Doing it Better*. Oxford: Oxford University Press.
- Douglas, H. (2009). *Science, Policy and the Value-Free Ideal*. University of Pittsburgh Press.
- Dupre, J. (2020). "Following the science" in the COVID-19 pandemic. *Nuffield Council of Bioethics Blog*. <https://www.nuffieldbioethics.org/blog/following-the-science-in-the-covid-19-pandemic>
- European Commission (2020). *Digital Services Act Package*. <https://ec.europa.eu/digital-single-market/en/digital-services-act-package>
- Foucault, M. (2003 [1963]). *The Birth of the Clinic*. Routledge Classics, 3rd English edition.
- Krige, J. & Leonelli, S. (2021) Mobilizing the Translational History of Knowledge Flows: COVID-19 and the Politics of Knowledge at the Borders. *History and Technology*.
- Jee, C. (2020). Is a successful contact tracing app possible? *MIT Technology Review*. <https://www.technologyreview.com/2020/08/10/1006174/covid-contract-tracing-app-germany-ireland-success/>
- Leonelli, S. (2019) *La Recherche Scientifique à l'Ère des Big Data: Cinq Façons Donc les Données Massive Nuisent à la Science, et Comment la Sauver*. Éditions Mimésis.
- Research Data Alliance (2020). *RDA COVID-19 Working Group*. <https://www.rd-alliance.org/groups/rda-covid19>

This article is © 2021 by the author(s). The editorial is licensed under a Creative Commons Attribution (CC BY 4.0) International license (<https://creativecommons.org/licenses/by/4.0/legalcode>), except where otherwise indicated with respect to particular material included in the article. The article should be attributed to the authors identified above.