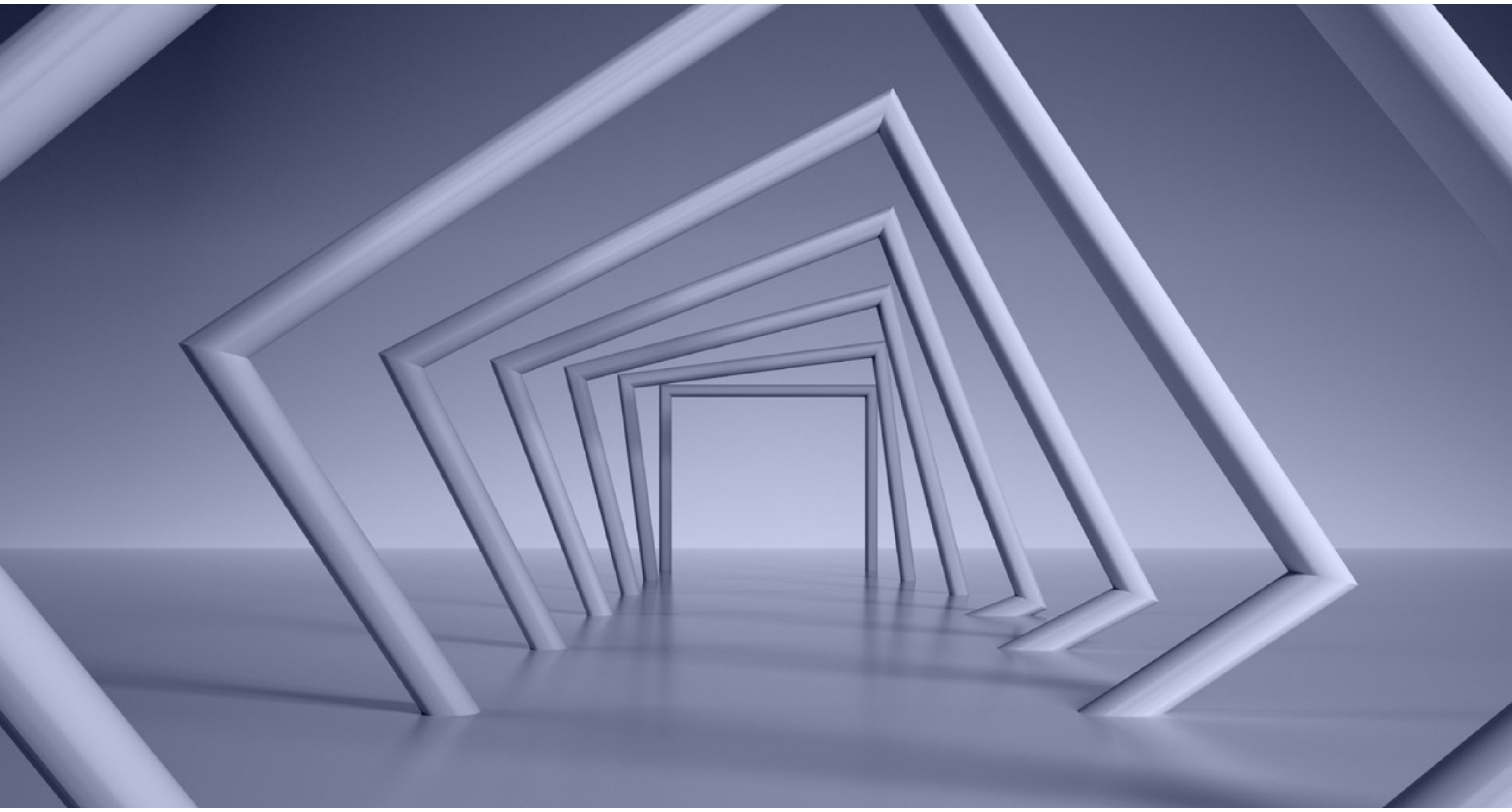
McKinsey
Analytics

How to make the most of AI? Open up and share data

Open data can unlock solutions to pressing challenges, says the Open Data Institute’s Jeni Tennison. But data quality, security, and privacy need to be ensured.



© Getty Images

June 2020

In this episode of the *McKinsey on AI* podcast miniseries, McKinsey's David DeLallo and Jeni Tennison, vice president and chief strategy adviser at the Open Data Institute (ODI), discuss the promise and pitfalls of open data. They explore ways that applications of open data can benefit society and consider efforts to preserve data security and privacy, promote data discovery, and ensure the quality of open data.

Podcast transcript

David DeLallo: Data—it's the lifeblood of the AI techniques used most often today. Most organizations have plenty of data within their veritable walls to fuel AI applications that improve areas from operations to product offerings. But it's the sharing of data across organizations that could unlock huge benefits for society. There's the potential to find cures to disease, to respond more effectively to crises, to combat climate change.

Today, however, very little of the data sharing needed for such endeavors is happening. And this is due to a variety of reasons, from technical challenges to very legitimate privacy concerns and also because many organizations are simply hesitant to share their data because they see it as providing them with competitive advantage.

So how can we break down some of these barriers to unlock the potential for data and AI to do societal good? I'm David DeLallo with McKinsey Publishing, and I explored this question in a fascinating conversation I had with vice president and chief strategy adviser at the Open Data Institute, Jeni Tennison. The London-based institute works with businesses, governments, and other organizations to build an open and trustworthy data ecosystem.

The first order of business I wanted to tackle with Jeni was to understand exactly what she and the institute mean when they say "open data." Here's how she explained it.

Jeni Tennison: We think about data as being on a spectrum of access. So there's data that's closed within a particular organization—only that organization and people within that organization can access it. And at the other end of the spectrum, there's open data, which is data that anyone can access, use for whatever purpose, and share with others.

Fully open data tends to be data that comes from organizations that care about getting maximum value for the economy and society as a whole from that data. And it tends to be data that is not personal data.

A good example is lidar data, which shows you hills and valleys and so on. It can be used to predict flooding. That data is made available as open data in the United Kingdom, so that organizations can use it to create accurate insurance premiums, for example.

But you can see that that's very nonpersonal data. So open data really spans a wide range of different types of data, tends to be nonpersonal but doesn't have to be nonpersonal, and is made available because we think that others can benefit from it and that we actually get value from others benefiting from it.

David DeLallo: Of course, it's that personal data that can be leveraged best for ways that would benefit society as a whole—for example, by allowing doctors to understand the nature of disease and possibly see a path to new treatments. Anonymization and aggregation techniques are one way such data can be shared and leveraged. I asked Jeni if there were other ways to go about safely using personal data. And she talked about an idea that's being explored more recently, and that's the concept of a data trust.

Jeni Tennison: One thing that we [ODI] have been looking at is this concept of data trusts as third-party institutions that would wrap within them the kind of governance that you need to make sure that the data is being shared well and for good purposes, for the purpose that it's going to benefit the people who are affected by that data.

So the idea of a data trust is that it would be an institution set up with a particular purpose in mind. The people running that institution, the trustees, would have a responsibility and a legal responsibility to share data only in accordance with that purpose of the trust.

And then you would obviously need to have the kind of technology that makes that data easy to share. You would have to have very strong engagement with the people who were affected by the sharing of that data in order to make sure that there is trust in the way it's being shared for that purpose.

But setting up those kinds of institutions is a direction I think we will travel in as we try to get more use out of data, both for public-good purposes and for understanding how it can unlock innovation and economic growth.

David DeLallo: But while data trusts might solve issues around privacy and security, they don't address the concern from organizations that they'd be giving up some competitive advantage by sharing their valuable data. Jeni pointed out, however, that there are plenty of reasons to share data that would unlock more value than keeping it close to the vest.

Jeni Tennison: When we look at the reasons that private-sector organizations start to share data, they fall into a number of categories. Sometimes private-sector organizations do things for social-peer purposes. And you can be cynical and just view that as them trying to build up their reputation or those kinds of things. But some organizations actually have at the heart of them that they want to do good. So there's that kind of category.

There's another category of sharing and opening data that is really about driving open innovation. So many organizations recognize—particularly larger ones—that being a bit more porous in the way in which they interact with other organizations out there in their environment can bring benefits back to them.

We have a project called Data Pitch, which is funded by the European Commission, where we're kind of matchmaking organizations that have lots of data with small start-up AI businesses. And the organizations that have a lot of data have a lot of challenges, too, that they think that the data they have can help with but don't have the know-how themselves about how to do that. They are sharing data with those smaller organizations in order to get that innovation back into their organization themselves.

In other places, we have organizations that recognize that they are part of a much wider ecosystem of organizations that is under a significant challenge. For example, we do quite a lot of work in the agriculture sector. The agriculture sector recognizes that with growing populations and with climate change, it is under an increased threat and demand around being able to satisfy the need that we will have in the future. And the only way in which that can be addressed is by organizations like Syngenta and Monsanto actually working together, and sharing more data is a way of doing that to ensure that we can feed the world in the future. So there are those kinds of larger challenges that get addressed through more sharing of data.

David DeLallo: And many organizations that collect the largest volumes of personal data—think big tech here—still have elements of their data they could share.

Jeni Tennison: First is open data that's generated from that personal data. So an example of that would be Uber, which aggregates data about journeys made through cities by its drivers and then uses that data to help inform the way in which cities might put in other traffic restrictions or can then predict what the impact of a particular event might be on the way in which traffic moves around the city.

Now, Uber's data isn't quite open data. The company doesn't make it available for anyone to access, use, and share. It's not available for commercial purposes. But it's the kind of aggregate data that is nonpersonal but could be open for anybody to use.

The second way in which organizations like that are looking at sharing data more is providing restricted access to bulk data about the people who are using their services but for public-good research purposes. So an example of that would be, say, Facebook making available data about the social graph in Facebook and the way that people use Facebook, so that social researchers can understand more about the way in which Facebook works but also how we interact with each other as people, which is valuable research for the future.

And then the third way in which those kinds of organizations are making data more available, more open, is by enabling us as individuals to get access to data that is about us and enabling us as individuals to choose to pass that data on to other organizations and other services. In the United Kingdom, we have that as a right. We have the right of data portability through GDPR [General Data Protection Regulation]. We have the right to get hold of data that is collected about us, pass that on to third-party organizations, to other applications, that can then use it to provide insight about us.

A good example of that is the open-banking work that has been happening in the United Kingdom over the past couple of years, which gives people the ability to access information about their bank transactions and their bank accounts and then to pass that on to third parties, which might be another bank if they wanted to move banks.

But more often, it is actually insight-type applications that are giving them a better understanding about the way in which their bank balance flows up and down and therefore helping them to manage debt, helping them to manage when they might need to get loans, for example, for a small business. So you can see what more openness and more sharing of those three sweet spots [can enable].

David DeLallo: Even for those organizations willing to share data, there's another obstacle: the physical or technical aspects of sharing. Jeni talked about some ways to tackle those.

Jeni Tennison: There's really interesting interplay between the technology that you use when sharing data and the governance that you need to have when sharing data. The bit that I'm most interested in is the use of new privacy-enhancing technologies that enable us to have more sharing without invading people's privacy—things like really good, high-quality synthetic-data generation, where synthetic data (so it's just made up) can be shared with organizations, so they can try out their algorithms and their data processing before getting hold of the real thing. That, I think, helps to reduce some of the costs and barriers that you would otherwise have when developing algorithms or getting access to data.

The other thing that I'm very interested in and excited about is the new advances around kind of distributed machine learning, where individual organizations can still keep hold of data themselves—it's still behind their own kind of firewalls—but the machine learning can happen on that data. And then the results of it come back into the central place and are combined for multiple different organizations, so that the person who is creating the AI never gets behind those firewalls, never gets access to that personal data, but we're still able to learn from it.

Those are the kinds of technologies that we need to explore more, because the more we can get away with not sharing huge amounts of bulk data that's very, very personal, the better the situation will be.

David DeLallo: Interestingly, Jeni pointed out that one of the biggest barriers to data sharing, even within organizations, is data discovery. And while this is often viewed as a technological challenge, it's actually one that's best solved by humans.

Jeni Tennison: So how do you find out that some other person in your organization—or some other organization, if you're looking more widely—actually has some data that you think is going to be worthwhile for the challenge that you're currently facing?

One of the things we find with discovery of data is that the curatorial role of the librarian is closer to what we need than a technology solution. It's the individual who can actively do the research to find out what data is available and pull that into something that makes sense, that tells the story of what data is available, that helps to articulate to people the subtleties about the quality of data, its provenance, what kinds of conclusions it can support, and what kinds it can't support. These are subtle kinds of notions that need to be articulated by people rather than technologies.

David DeLallo: While we were on the topic of data discovery, I asked Jeni if she had some tips for organizations around how to assess the quality of data that's available to be shared. After all, in research we conducted last year, more than a third of senior executives admitted that the data their organizations use to make the most critical decisions is of a somewhat low quality.

Jeni Tennison: Data quality is a really tricky issue. I think one thing to be very mindful of is that quality is associated with use. It's quality for use rather than quality per se. So the same data might be of completely adequate quality for one set of conclusions and completely dreadful for another set of conclusions.

So that's when we come down to this need to really, really understand the problem that you want to apply that data to in order to understand whether the quality of the data that you're getting can match that use.

Now, there are some bits about quality that are very obvious, very standard. Things like completeness, missing values, standardization, or lack of standardization are the kinds of things that can be assessed by tools, that you can get scores on. And that, of course, can help inform a decision about whether that data is fit for purpose or how much work you're going to have to do in order to get it fit for purpose for the work that you want to do.

But much more important when it comes to quality is where it comes from, how it was collected, what kinds of biases there may have been in that collection, what kind of error rates there are. That's much more qualitative information.

That information can be extremely hard for a third party who's assessing data to actually get hold of. And this, I think, is where we come to the bit that is really one of our biggest challenges around sharing of data: getting the originator of data to describe all of that background information about how the data came into being, which can enable really informed choices about the degree to which it can be trusted, the degree to which you can draw particular conclusions of the backfit.

Getting organizations to articulate that is very, very hard. We've been trained to think of the kind of metadata that we need to supply around data sets as being very factual kinds of things. When was it created? Who is responsible for it?

We need to really be learning from statisticians and the way in which they provide the full methodology—real details from statisticians, from scientists, about how they think of describing the real detail of where data has arisen from—so that we can get into a state where it's actually possible for third parties to assess whether data can be used for the particular purpose they want to use it for.

David DeLallo: And with that final insight from my interesting conversation with Jeni, we bring this episode to a close. We hope you enjoyed listening and that you'll check out the other podcasts we have available for you on this and other McKinsey podcast channels. Thanks so much for joining.

David DeLallo is an executive editor in McKinsey Publishing, based in McKinsey's Stamford office. **Jeni Tennison** is vice president and chief strategy adviser at the Open Data Institute.