"Coronavirus is showing that we need to understand data better"

Statistician Katharina Schüller is part of the team of the popular science campaign 'Unstatistics of the Month' - a group that critically scrutinises how figures are presented in public. In this interview, she explains why we often unconsciously classify data incorrectly.

© Studio GOOD

Ms. Schüller, for more than a year and a half we have been literally "bombarded" with projections on coronavirus. Particularly at the beginning, forecasts about the spread of the pandemic often proved to be inaccurate. Why was that?
The assessment of a risk to support further decision-making is always based on data that already exists. But precisely these data are not particularly good for assessing the situation, especially at the beginning of a crisis because that is not why they were collated, and so may not be representative, for example. We cannot derive any precise recommended actions from them. As best, they can serve as a guideline for our actions.

Were we too certain too soon?
The forecasts often gave that impression. But there are two sources of uncertainty we need to consider. One lies in the nature of any data analysis: estimates are always inaccurate to some degree. These uncertainties can be expressed as ranges of variation, or 'confidence intervals'. Furthermore, there can always be influences on what is happening that are not foreseeable, such as virus mutations, fluctuations in the weather or reactions to the projection. This happened during the coronavirus pandemic.

What lessons can you draw from this?
Coronavirus is showing us the importance of solid information and a robust data infrastructure. What data is available is also crucial. There's a saying that 'You can't manage what you can't measure'. It means that where information is lacking, there are blind spots in the management of a crisis as well as problems you fail to see. There is also a cognitive bias.

What does that mean?
We overvalue incomplete data. Simply because they are there.

99

We overvalue incomplete data. Simply because they are there.

© privat

Katharina Schüller encourages a conscious approach to statistics. She is on the board of the German Statistical Society and director of the Munich-based consultancy 'Stat-up'.

Could you give us an example?

If the current number of coronavirus cases reported in the headlines every day creeps up and down, then at some point it is about this one factor alone. At the same time, we ignore gaps in knowledge. Because what would happen if the 'Tagesschau' news programme were also to report every day on how many people have lost their job because of the pandemic? Or how many school lessons were cancelled? Or how often depression was diagnosed? That would give us a very different picture of the consequences of the pandemic.

So it's wrong to concentrate solely on the infection figures?

Yes, absolutely. You need to access the most diverse data sources to get a comprehensive picture of a crisis. These must be assessed and viewed in relation to each other. After all, millions of people with very different circumstances are affected by the coronavirus crisis. Single

mothers with children of school age, for example, who often have a hard time. We need to take this diversity of perspectives seriously. We will otherwise not recognise that there may be conflicting aims and measures that could have undesirable side-effects.

How can we take better account of this diversity of perspectives?

The question is: which aspects of reality do I want to include? We should understand a pandemic like coronavirus as a complex system: what points of view are relevant to navigate this system and to come out of the pandemic well? Which goals are important? Not all data are helpful. But expert analysis of these data is a key factor in drawing consequences from such a situation and controlling it – and making a lot of what we experienced in the past year more manageable.

One focus of your work is Data Literacy – the expert handling of data. Can Data Literacy improve our ability to assess risks?

For me, handling data is essential for risk competence. We need to learn to scrutinise data and information critically. We need to understand what is in the data, and what is only added as a result of our interpretation. Assessment is never purely objective, it also depends on the goals being pursued. With coronavirus, for example, the question is: are the measures that are taken solely about averting an immediate threat, or are we interested in medium and long-term consequences and issues around quality of life? Depending on our aims, we need to classify data according to specific criteria and assess them accordingly.

What practical consequences should we draw from that?

The decision-makers, politicians for example, need a better understanding of data: what are the strengths of data, where are the limitations and what are the opportunities? They need to know how to communicate data – including the unknowns, which we always have to think about too. We also need high quality public data and statistics based on a dependable infrastructure. That's one thing that is often forgotten when we talk about expertise in a crisis. Building a data infrastructure may not sound as hip and sexy as keywords like Big Data, Artificial Intelligence or Dashboards – but we need a quality-assured, professional system for data provision and analysis that policymakers and administrators can access reliably.

How moral is it to collect data? The key term is data ethics.

When it comes to collecting and processing information, the moral question is very often only asked in terms of what is not allowed. As if data ethics is only about not misusing data. But data ethics is also about what you should do: use data for a good purpose and to the benefit of society. Because not using data even though they could help to solve problems like the current pandemic better and faster is just as unethical as misuse. ◨