Computer Science Track

Interview with

Jaap-Henk Hoepman

Jaap-Henk Hoepman is co-curator (with Marit Hansen) of this year's CPDP Computer Science Track. He is an associate professor at the Digital Security group of the Radboud University, Nijmegen, and a principal scientist (and former scientific director and co-founder) of the Privacy & Identity Lab. In October 2021 his book Privacy Is Hard and Seven Other Myths. Achieving Privacy through Careful Design appeared at MIT Press.

You are known to many people in the CPDP community, but could you share a few words about yourself, your work, your relationship to CPDP?

I think I'm a long time participant at the CPDP, to be honest. I don't even remember when my first time was. It might easily have been 15 years ago, if that's how long CPDP exists. I don't know. And I've always gone because it's such a nice mix of policy people, lawyers and also a little bit of computer scientists, but not so many – all of which work on data protection and privacy from all across the world. And yeah, for me that's super interesting because I'm a computer scientist working in privacy, privacy by design, privacy enhancing technologies. I like to work on problems that are inspired by practice. This is a good source of new problems, new issues.

For this year's edition, CPDP is grateful that you are co-curating the computer science track. How do you see this and what would be a good outcome of this track?

We don't have a lot of experience yet. I mean, it's like an experiment. The idea was to maybe attract more computer scientists to CPDP because there's few of them compared to other "factions" or how you want to call them. There are not many and it would be good if there are more. Let's see if that happens. So to say whether the experiment was successful or not is something that I can only answer after CPDP, I guess. For me, the main thing is seeing more fellow computer scientists, then it will have been good.

Why do you think CPDP is a good platform for computer scientists to get together?

Well, CPDP will be an awful platform for computer scientists to get together with only computer scientists. But that's not the point. The point is that they get together with policy people, lawmakers, legal scholars, … That's the whole point. And CPDP is a very good venue for that. I've always enjoyed it, and I think for many computer scientists it's relatively unknown, and therefore we are trying this to get them involved.

On a general level, how do computer scientists benefit from more interaction with legal and policy experts?

On several levels. I think in general there is a significant number of computer scientists working in data protection and privacy who don't necessarily have a very good understanding of the more social and the more legal aspects of it, which means that if they think about solutions from their discipline, they may sometimes solve the wrong problem. Or they may solve a problem in the wrong way, because they're not really accounting for the social and legal dimension. So that is one reason why it is important, and the other thing is that, sometimes from a computer science perspective, there's not a problem at all, but then, from a legal perspective, there is. So, you might think as a computer scientist that you've completely solved the problem, but it's not solved. This, in general, depends a bit on the kind of computer scientist that you are, but if you want to make your work more relevant, it is nice to get inspired by actual problems that occur in practice. There are many ways that you can get inspired by problems that happen in practice. You might work with companies, you might engage with civil society, or you can go through other disciplines to see how they look at this problem and get inspired.

Let us turn to a technical topic – the local processing of personal data, which is becoming quite important in the age of AI. As you mentioned in your book ("Privacy is Hard and Seven Other Myths"), iPhones have been doing it, and now Microsoft Copilot is doing it. How do you feel about this way of processing personal data? Does it provide adequate privacy guarantees?

Local processing in general is a way to protect privacy because it means indeed that your specific data does not leave your own device and therefore does not risk being abused by any third party, which is of course one of the reasons why we want to have privacy and data protection.

The example I give in my book is the example of Apple allowing you to search only on your device for pictures which happen to have a certain face. In that case, that specific application is clearly for your own benefit. You want to have photo albums categorised based on the people on the pictures and that's it. But more and more of the local processing is done through

> In general there is a significant number of computer scientists working in data protection and privacy who don't necessarily have a very good understanding of the more social and the more legal aspects of it, which means that if they think about solutions from their discipline, they may sometimes solve the wrong problem. Or they may solve a problem in the wrong way, because they're not really accounting for the social and legal dimension.

things called federated learning. What the system tries to do then is to observe your local behaviour, to derive certain information about you and to collectively construct a model about, let's say, the typical behaviour of Northwestern Europeans regarding the use of Signal. You can do all kinds of trainings or [find out] which kind of words or emojis are most popular. That kind of stuff. And so that data is then fed into a model that is then actually used by those companies for all kinds of purposes that might be harmful to us as individuals or to society. Then you would say, strictly speaking, privacy is protected and there's this concept of differential privacy where mathematically speaking (and you can have debates on how good the protection actually is), but the idea is that the system that you built, and the analysis that you get out of that system, is such that you can actually prove that it doesn't matter that your personal data is in or is out of that dataset being used to answer the question. That means that, strictly speaking, your privacy is protected because it doesn't matter that you are being interviewed or not, but again, collective inferences can be made that have an impact on society and have an impact on you individually.

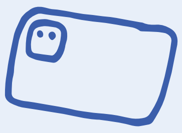
The last paragraph of your book says: "Significantly improving the privacy of the apps and services we use should be our first priority. But at some point, we need to dig deeper down into the technology stack and reconsider the designs for our computers and networks, both at the hardware and the operating system levels. These designs are half a century old by now and never fundamentally changed, while the world in which they are used has changed beyond recognition. We are stretching the boundaries of their use beyond the breaking point – not only in terms of privacy but also in terms of security and reliability. It's time to start redoing the plumbing, instead of applying Band-Aids to temporarily stop some leakage while we frantically mop the floor against all odds." The world has continued to change in the few years since you wrote these words – has technology and privacy by design caught up a bit, or started to catch up?

Perhaps a little bit at the highest layers of the stacks, in terms of things like federated learning, but less if you look at the lower layers of the stack. Take for example the results of a seminar I was at last week, where people were presenting about how home automation works. You have these home automation devices, more and more of which get introduced into our homes – light bulbs, voice assistants, TVs, everything is smart these days. It turns out that these devices are programmed in such a way that they constantly communicate to get, for instance, in touch with devices that are useful to them, so they can e.g. provide services together. This leaves lots of traces that you could pick up if you wanted to. Now it turns out (and I was surprised about this), that there's quite a few software development kits that are embedded in software that people write that can actively try to obtain that information. That basically means that, if I'm using a phone and I would have some of these applications with these software development kits installed on my phone, then my phone will be spying on my home to see what kind of Internet of Things devices are in my home, or if I'm somewhere else, what kind of IoT devices are in the other place. So, it turns out that things like browser fingerprinting, which was restricted to browsers, which is bad enough, is now sort of expanding to physical fingerprinting – what devices are in this space? And of course, everybody has a different set of devices, which is a unique fingerprint, and it tells a lot about the kinds of stuff that you have and do. In that sense it's getting worse, because more and more of this technology gets sort of embedded into our daily lives, and this was the promise of the ubiquitous computing, the computing disappearing in the background. But it is not disappearing for anybody who is listening to all the signals and things that they are doing to do so. And in that sense, it is not getting better at all.

Can you tell us a bit about the projects that you are working on now?

I have been quite busy with the whole European identity wallet developments as eIDAS2.0 got accepted. This is a nice topic for CPDP. Sometimes I jokingly say that I am a computer scientist that reads regulations for breakfast. It is important as a computer scientist to read these things and to engage with them because if you read the regulations, they are full of nice promises of how important privacy is. Super, that's excellent, but the actual privacy properties of something like the European identity wallet are only defined by the actual implementation of these wallets, and the problem is that you cannot in any legal language describe how such a system should work to ensure that it satisfies the properties that you want. The actual regulation tries very hard to do that, but it fails, and it's not the fault of anybody, it's just that you cannot do it. Legal language is not precise enough to do that. This means that you must engage with the thing that defines and specifies the European identity wallet and that has a very strong impact on the security, privacy, trustworthiness of the overall system (in this case it's called the Architecture Reference Framework). And that means that the writing, the development of that specification should be given the same political deliberations as the legal texts do, and I know for a fact that certain members of parliament at the time expressed concerns about the fact that they could only debate the regulation text, but not the technical specifications. Now whether they are competent enough to discuss and decide the technical documents is another discussion, but you see there that it is important to also have computer scientists, but also civil society more onboard when these technical standards are being challenged. CPDP is a nice platform to talk about these things, why they are important, but the incentives are not very well aligned. It is hard to get a computer scientist, or civil society on board because it takes a lot of time. Civil society doesn't have a lot of money; computer scientists maybe do but I mean they are academics in general, we don't get rewarded for that kind of work.

Lately, I'm thinking more about the question of the difference between the digital and the old physical world in terms of disruptive power of the (let's say) "non-discreteness" of the digital world, meaning that in the digital world, you may spend a lot of money developing a service but once it is developed, you can scale it almost for free. And in that sense, maybe it's relevant to the theme "The World is Watching" because this has tremendous impact on surveillance powers, so states can invest research in order to develop surveillance tools and once the tool for surveillance is out there, it scales to millions, billions of people, whereas in the traditional world, if you wanted to surveil people, you would have to have real human agents to do that. If you want to surveil 10 people day and night, you will have to have 30 agents to do the surveillance in shifts. At some point you must decide because you don't have that many agents. You must decide, I'm going to watch that guy, and not that person, because you don't have that many agents out there. In the digital world it doesn't work that way, you can very easily with the same amount of effort needed surveil 10 people or surveil a billion people. The effort is roughly the same. And this is a concern, it makes the whole debate about privacy vs. security very binary in the sense that either you have full privacy, or you have full surveillance, there is nothing in between. It makes the discussion very complex. I am trying to think of ways that you can add friction to surveillance tools so that you can guarantee a tool never scales to surveillance of a million people, because you would have to spend a lot of actual physical resources to do that. Similar things could also be very interesting to study, like the spread of fake news, or the first mover advantage, where the first person to have a big share of a specific market is the one that's going to dominate the market, who is going to have the monopoly, Because nobody else can actually enter the market anymore. Maybe friction there can also help us. This is a very speculative area of research that I now only started to think about. ▪



The problem is that you cannot in any legal language describe how such a system should work to ensure that it satisfies the properties that you want… And that means that the writing, the development of that specification should be given the same political deliberations as the legal texts do…

Computer Science and Data Protection Officer Tracks

As a novelty at CPDP.ai, this year's edition features two "dedicated tracks". Join us on Wednesday to engage in multidisciplinary debates around computer science topics. On Thursday, data protection officers and other privacy professionals get together to discuss subjects of particular interest to data protection officers, and the challenges and opportunities that these professionals face in an evolving digital landscape. Curated by experts with a strong grounding in these subjects, the dedicated tracks aim to nurture interdisciplinary discussions and connections amongst specialists and with the wider CPDP community.