AI ETHICS

Algorithmic Transparency in the Public Sector

By Natalia Domagala, Head of Data Ethics at the UK Cabinet Office

In this module I draw on my professional experience working on data ethics, open data and open government to explain algorithmic transparency in the public sector. I begin by defining the concept and explaining why it is critical in our society today. I briefly review different examples of algorithmic transparency measures from Europe and North America before offering a detailed explanation of the UK's new Algorithmic Transparency Standard. I conclude with my outlook on the field of algorithmic transparency in the next few years and a few suggestions on what actors in the field ought to focus on going forward.

Lecture Transcript

0:02 Welcome, everyone. It's a real pleasure to have this opportunity to talk about algorithmic transparency with you as a part of the AI Ethics:Global Perspectives course. Thank you very much to the NYU GovLab team for inviting me to be a part of this fantastic initiative. My name is Natalia Domagala. And I currently work as the Head of Data Ethics at the Central Digital and Data Office, a part of the Cabinet Office in the UK Government. Previously, I worked on open data and open government for the UK government and for civil society organizations. Today, I will cover algorithmic transparency, which has been the main focus of my professional efforts. And I will begin by defining algorithmic transparency and explaining why it matters. I will then show you various existing examples of algorithmic transparency measures. I will give you more details about the algorithmic transparency work that I've been leading in the UK. And finally, I will offer a few suggestions on what I think the focus of this field should be in the next couple of years. And before I begin, I'd like to stress that this talk is focused on algorithmic transparency in the public sector, mainly due to my expertise and my experience, but I'm well aware that similar efforts exist in the private sector and are very advanced and impressive as well. And although I'll be drawing on my work and describing the development of algorithmic transparency standards in the UK, the views and opinions I will share in this lecture are personal and do not necessarily represent those of my organization unless explicitly stated. So let's begin. There is a person at the end of every policy decision that we make as the government. Many of these decisions are automated, and this is for good reasons, such as improving efficiency or saving 1000s of pounds of taxpayers money. And some of those automated decisions can affect individuals in a life changing way. For example, by determining their provision of essential public services. A person whose application for financial support has been rejected might be left wondering why this happened, and how exactly the government made the decision that affected their life. Algorithmic transparency is all about communicating clearly how these decisions are reached, and what role algorithmic tools play in this process. Now, algorithmic transparency matters because the public has a democratic right to information and explanation about how the government operates and how it makes decisions. In the UK, the public also has data rights under the UK GDPR. In Europe, that's under GDPR. And they're also local equivalents of that legal framework in other countries. Opening algorithmic assisted decisions up to scrutiny helps

build and maintain trust in government. Algorithmic transparency also provides an opportunity for government departments to highlight good practice around the use of algorithmic tools. It facilitates learning both within and across government departments. It contributes to improvements in the development, design and deployment of algorithmic tools across the broader public sector. It enables those who build, deploy, use or regulate algorithmic tools to identify any potential problems with a given tool early on, and potentially limit future negative impacts. And for external suppliers, it helps demonstrate their commitment to ethical service delivery, showcase their value proposition, and improve public trust in their systems. Now, there are many algorithmic transparency initiatives that already exist around the world. And I would like to show you just a few. These are the main ones that I'm aware of. I know that there must be many other algorithmic transparency initiatives that I don't cover in this lecture. So please don't think that this is the exhaustive list of what's happening in terms of algorithmic transparency around the world. So let's begin with an example from New York City. In September 2020, the Algorithms Management and Policy Officer in New York City launched the first ever agency compliance reporting process, which essentially asked New York city agencies to provide information on level one algorithmic tools in use, and as required by Executive Order 50, the compliance process resulted in a PDF directory of high priority algorithmic tools currently in use by city agencies. And as you can see, the report has information on what agency is using the tool, what's the name of the tool, when the tool entered usage, what's the purpose of the tool and what's the overall function. And in the report, you can also find definitions of the algorithmic tool and what it means to be a high priority algorithmic tool.

5:02 Another example comes from Canada. So, in Canada, there is a tool called Algorithmic Impact Assessment, which is essentially for public sector departments using algorithmic tools. And as a part of this assessment process, information is gathered on what kind of impact the tool has on individuals and entities, what the impact is on government institutions, what are the data management processes, what are the procedural fairness considerations, and what's the complexity. And the final results of this algorithmic impact assessment can be published on the Open Canada website. However, there isn't currently a centralized database compiling all the results. In France, algorithmic transparency measures currently include those categories of information. So we have an overview of the administration. So the name of the administration where the algorithm's deployed, contact details, information update date. We then have business information on the algorithm and the decision taken such as the global context, name, purpose, how and when the algorithm intervenes in decision making processes, levels of decision automation, legal foundations and related resources. We have information on the impact of the decision, including the number of the decisions taken per year, scope of the decision, public affected by the decision, and related resources. And information on the internal workings of algorithms. So all the technical specifications on the algorithm and the data. And then finally, we've got a Helsinki and Amsterdam example, which is possibly the most well, the most famous one. And as you can see here, that's AI Register website from Helsinki. So it covers all the examples of these, of AI in the City of Helsinki. It's available in English and Finnish. And each of these examples are described in those categories. So we have information on the datasets, data processing, non discrimination, human oversight, and risks. And the last example I'd like to cover is from the Netherlands. And in the Netherlands, we have an algorithmic transparency standard also available in Dutch and English. And we have information such as the name, organization, department, short description of the algorithm, type of algorithm, any URLs if there is a separate website for that, status, goals of the policy or or the model, impact, proportionality considerations, details in the decision making process, any additional documentation, source data methods and models that the algorithm is using, monitoring human intervention risks,

performance standards, any lawful basis, data protection impact assessment and objection procedure, as well as the revision dates of the of those entries. So now, I would like to give you more details about how algorithmic transparency looks in the UK. And I would like to begin by emphasizing that the need for greater transparency on the use of algorithms in the public sector in the UK and the need to standardize this information have been continuously flagged and recommended by a number of bodies and organizations. And the full list is much longer than what you can see here. But I would like to bring up just a few examples of those recommendations.

8:39 So firstly, recommendations from the Committee on Standards in Public Life said that the government should establish guidelines for public bodies about the disclosure of their AI systems. A think tank called Reform recommended that public sector organizations should list and publish their use of algorithms, their operating logic and their governance arrangements on a dedicated search book of the UK page. The Center for Data Ethics and Innovation bias review said that the government should place a mandatory transparency obligation on all public sector organizations using algorithms that have a significant influence on decisions affecting individuals. In their "Understanding Artificial Intelligence, Ethics and Safety", The Alan Turing Institute flag that transparency of AI should demonstrate that both the design and implementation processes that have gone into a particular decision of the system and the decision itself are ethically permissible, non discriminatory and worthy of public trust. And finally Ada Lovelace Institute, AI Now Institute and Open Government Partnership, in their study on algorithmic accountability for the public sector, stressed that meaningful transparency is often the prerequisite for increasing algorithmic accountability. Now, let me explain the process of developing the standards that we've worked on collaboratively with the Center for Data Ethics and Innovation. So, we began with a series of workshops with external experts. And during those workshops, we asked them what information on the use of algorithms should be included in the algorithmic transparency model? How should this information be presented, how often it should be updated and which algorithms should be in scope? And we then consulted these findings with our colleagues from across the government and asked them similar questions. Finally, again, jointly with the Center for Data Ethics and Innovation, we commissioned deliberative research with an aim to consider how the public sector can be meaningfully transparent about algorithmic decision making. The core objectives were to explore which algorithmic transparency measures would be most effective at increasing public trust, and public understanding about these algorithms in the public sector. And for this study, we recruited participants from all over the UK. And we would we concluded that whole process in the codesign session, when we essentially asked them, how, what kind of algorithmic transparency measures they would like to see, and we let them design those measures themselves. And there are quite a few recommendations coming from this study that we've directly implemented. But I will cover that in more detail later when I describe the UK Algorithmic Transparency Standard. So this is the list of categories identified in the workshops. And this was the very foundation of our standard. And before I show you the standards, I'd like to specify what we mean by an algorithmic tool and also what's our definition of an algorithm. So an algorithm is a set of step by step instructions in artificial intelligence. The algorithm tells the machine how to go about finding answers to a question or solutions to a problem. And an algorithmic tool is a deliberately broad term that covers different applications of AI and complex algorithms in a consistent manner understandable to non expert audiences, and by algorithmic tool we mean a product application or device which has been deployed to support or solve a specific problem using complex algorithms. And this tool might have been developed in-house or bought from a third party.

12:12 So based on all these research exercises, we developed the Algorithm Transparency Standards to help public sector bodies in the UK share information on their use of algorithmic tools with the general public. And the public sector bodies can provide this information by filling out the set template and publishing it on gov.uk. In terms of applicability, in the initial phase of this work, we will prioritize tools that either engage directly with the public such as chatbots, or tools that meet at least one criteria in each of the three areas. So the first area is technical specifications. So complex statistical analysis, complex data analytics, or machine learning could be neural nets, deep learning, or any other forms of machine learning. The second category is a potential public effect. So tools that have a potential legal, economic, a similar impact on individuals or populations, or that affects procedural or substantive rights, or eligibility received or denial of a program. And the third category is impact on decisions. So tools that replace human decision making, and assess or add to human decision making, for example, provide evidence for decisions. And this prioritization system that we developed is a very, is a very first attempt to actually make sense of that. And this is something that will keep reiterating and updating and I think it's really important to stress that understanding what tools should be in scope of algorithm transparency measures, is a huge challenge that algorithmic transparency policymakers and practitioners are tackling globally. Now, let me tell you more about the standard itself. It has two tiers. And this two tier system is a direct recommendation from the public engagement study that we run with the Center for Data Ethics and Innovation. So the members of the public that we ask in this study told us that they would prefer to have a layered level of information so that there should be a very short, very brief introduction to what this tool is and what it does. So tier one, and then tier two, with more detailed information for people who are really interested. So in tier one, you should give a very basic description about how the algorithmic tool functions, and why it was introduced into the decision process. And this includes things like what's the problem that the algorithmic tool is aiming to solve? And how the tool is solving that problem? What's the rationale for using it? It should be very short, concise, written in plain English, really easy to understand, with no technical vocabulary. And tier one is geared towards the general public so that interested individuals can learn about the tool and also know where and how to find out more information if they wanted to find more information. And tier two provides more detailed information about algorithmic tools. So this is directed at informed, slightly more informed and interested parties. That could be civil society organizations, journalists, people using algorithms and any other interested parties. And it has five categories. So owner and responsibility, description of the tool, information on the decision and human oversight, information on the data, and risk, mitigations and impact assessments. And I will now talk you through each of these categories very briefly. So in the owner and responsibility section, you should detail accountability for the deployment of the tool. So information such as what's the organization where the tool is used, was the team responsible, who's the senior responsible owner, were there any external suppliers or any third parties involved, what's the role of those external suppliers and what are the terms of their access to any government data. Then the second part is a description, a detailed description, of the tool, so we have more information about the scope. So what has this tool been designed for and what hasn't it been designed for. And here, we can also have a list of common misconceptions. And an explanation that this tool hasn't been intended for this particular purpose, which people who don't necessarily know much about algorithmic tools might assume. Then expanded justification section. So list of key benefits such as value for money, efficiency, ease for individual, and list of non algorithmic alternatives if they were considered in the process as well. And then we have a technical specification section, which is all about technical information, such as type of the model, how regularly the tool is used, what's the phase, what are the details of the maintenance schedule, and what's the system architecture. The next section covers the information on the decision and human oversight. So first integration, so what is the wider decision making process within the

tool operates, how is the algorithmic tool integrated into this process, what influence does the tool have on the decision making process. Then we have a section - human oversight, how much and what information does the tool provide to the decision maker, what kind of decisions humans take in this process, and is there any required training that people deploying this tool must undertake. And finally, a really important section on appeals and reviews. So what mechanisms are in place for view or appeal of the decision.

17:35 Then, we have a section on the data. So here we have all the information on datasets used to train the model, but also data sets that the model is or will be deployed on. And this includes information such as the name of the datasets, the URLs, if those datasets are openly available, information on the data collection process including the original purpose of data collection, information on the data sharing agreements in place, details on who has a will have access to this data, and how long to stay there stored for an under what circumstances. And the final section is on risks, mitigations and impact assessments. So the first part is a list of all the impact assessments that have been conducted with links and descriptions. And this can be assessments such as the data protection, impact assessments, equality impact assessment, any other algorithmic impact assessments, or data ethics assessments. And then in the risks and mitigations part, we have a detailed description of common risks for the tool, and the detailed description of actions taken to mitigate those risks. And we also have a short list of common risks, but we encourage teams to think about this really deep and provide an exhaustive list of all the risks that can possibly occur. And as this is a standard, we also develop the schema to support teams working through it and to make sure that the data on the use of algorithmic tools is consistent, and of high quality and this is something that we will keep developing further with the Data Standards Authority. Now, I would like to offer a few conclusions and perhaps provocations on what's next in this field in let's say, the next two years. So firstly, algorithmic transparency is still very new. It's an emerging area. And the existing mechanisms for public sector algorithmic transparency will need to be iterated as this field evolves. And in the UK, what we launched in November 2021 is the very first version of the standards and we're actively working on developing it further, based on the feedback from the public and organizations involved in the piloting process. And I think this approach of continuous feedback loops and iteration is something that other public sectors working with algorithmic transparency around the world will also have to take on board because this isn't something that's set in stone. We don't necessarily know what works best just yet, because it's such a new field, so we will have to do a bit of iterating and including feedback as we develop this algorithm transparency measures. The second point I would like to make is that there is a growing need for standardization and global alignment of algorithmic transparency mechanisms. And this is to ensure that they will be meaningfully comparable nationally and internationally. So let's say in five years, it would be ideal if we had enough information of similar quality, to be able to compare how algorithmic tools are being used across different public sectors around the world. And this is why we need to ask, how can we do that now, what kind of steps we can take right now, in order to make sure that this standardization happens in the next few months, and years. And another point that I think is really, really crucial is that public engagement is necessary when designing algorithmic transparency measures. A very key target audience of any algorithmic transparency efforts is the general public. So any measures should be developed to accommodate their needs and their preferences, and public engagement and deliberative exercises should be a starting point for designing and iterating any algorithmic transparency tools. And in the UK, we had that public engagement exercise that we ran with the Center for Data Ethics and Innovation. And this was invaluable in the whole process. And I think, as we move towards working more and more in the open and, and as deliberative exercises gain popularity across the globe as well, we should make sure that any algorithmic

transparency project will have those public engagement components as the very, very start of it. And finally, in order to be truly effective, algorithmic transparency needs to be accompanied by accountability measures. And I really can't stress this one enough. Transparency will not be sufficient if there is no accountability that follows. Thank you very much all for listening to this lecture. And good luck with the rest of your AI Ethics:Global Perspectives course. Thank you very much for having me.