Summary

- High levels principles that promote transparency, accountability, expertise and user-centred design will be critical for services that incorporate automated decision-making.

- This means that services which use automated decisions should be designed to:
  - Explain how decisions are made
  - Empower users if something isn’t right
  - Take context into account, acknowledging that ‘transparency’ and ‘accountability’ will need different design solutions for different users
  - Be interrogated by empowered and skilled experts trusted to audit the performance of automated decisions

- Projects by IF conducted a short research project to illustrate how these principles might apply in the context of a hypothetical benefit system.

- In this example, three groups of people are looking at a problem arising from the interplay between three algorithms.

1 Introduction

1.1 Projects by IF is a design studio that helps companies to build services that people trust. Our work has involved understanding the role algorithms play in people’s lives, exploring how services can be transparent, and designing patterns which help people understand services.

1.2 It is imperative that service providers are accountable for how they use data. This is crucial when it comes to services built around automated and semi-automated decisions.

1.3 We think the committee should focus its efforts on understanding the opportunities available to parliament to ensure that services which use algorithms:
  - Explain their decisions in terms that users can understand, and make it clear what is responsible for the decision (either an algorithm or a human supported by an algorithm)
  - Give users clear options when they are concerned about the decision an algorithm has made
  - Publish information that allows both effective auditing and support by third parties
  - Ensure that information about a service is published in a way that is accessible to experts

1.4 Something we’ve come to understand at IF is that transparency is more useful with context. To understand how something works, you have to give people the right information at the right time. The context that a service user needs is different to the context an auditor or a trusted third party needs.
1.5 To illustrate what we mean, we’ve worked up a hypothetical example for this submission. It breaks down how transparency could work in the context of a public service, and shows the interplay between different types of algorithms.

2 National Benefit Service: a hypothetical public service

2.1 Automated decision-making is already used in public services, for example in tax calculations. As more services are digitised, more automated decision-making will take place. It is also plausible that, in the next few years, government will begin to use machine learning to deliver parts of public services.

2.2 In our example system, we imagine three kinds of algorithm interoperating to deliver a service. This includes:

- A traditional algorithm to determine how much of a benefit a claimant is entitled to – this is how benefits are calculated today, based on conditions set by the government
- A machine learning algorithm to select appropriate appointment times and duration – this would use several data sources alongside historical information
- A semi-automated decision, where a traditional algorithm reports a claimant when they meet certain conditions but a human is responsible for applying a sanction

2.3 NOTE: The following examples are to illustrate the kind of interface and information necessary for making it clear how an algorithm has reached a decision. It should not be read as a vision for how a benefit service should be run.

3 Scenario 1: Showing the role of an algorithm to a user of a service

3.1 People using services should be able to see, in the context of the service itself, whether a decision has been made by a human or a machine. They shouldn’t have to go somewhere else to get that information.

3.2 It’s also important that a service gives people the opportunity to object to a decision if they think it’s based on incorrect data.
Written evidence submitted by Projects by IF (ALG0033)

3.3 When Mary’s benefit claim is paid, she is given two pieces of advice: how much she has been paid and the time of her next appointment.

3.4 When she goes to look at the payment amount, she sees that she has had a sanction applied. She can find out more from the service about why this has been applied and how the decision has been reached.

3.5 The important thing is that at each stage she can see if decisions have been made by a human, a machine, or a combination of both. She can also see how to take steps if something goes wrong.
3.6 In this example, we’re starting to see how several types of algorithm interoperate. Some obey strict rules described in government policy and law, while others are learning from a number of datasets at any one time. All of them need to be able to explain how a decision has been reached in language the claimant understands, particularly when those decisions are interdependent.

4 Scenario 2: Helping third party organisations understand decisions made by algorithms

4.1 There are limits to what individuals can do on their own. Organisations such as Citizens Advice provide critical support for people in distress. ‘Benefits’, ‘Work’ and ‘Debt and Money’ are the top categories on Citizens Advice website. Specific claims and caps frequently appear as the highest trending content.

4.2 When people come to these organisations for advice, they often bring large amounts of paperwork. Public services should be built in a way that allows people to share limited sets of data with trusted third parties so they can be supported when they’re in difficulty. In some instances, these organisations should also be able to run simulations of algorithms that make decisions.
4.3 After she misses several consecutive appointments, Mary's benefit payments are reduced. A friend in a similar situation suggests she speaks to someone at Citizens Advice for support.

4.4 Mary's case is handled by Subhit. As a Citizens Advice volunteer, Subhit has access to a service that is licensed to look into Mary's claim history and provide more detail about how some of the decisions have been reached. He's not able to see more than one case at a time, or any data about the machine learning algorithm involved in the service.

4.5 Subhit can see that Mary's circumstances and eligibility remain unchanged from previous appointments – only the missed appointments are flagged.

4.6 When Subhit explores more information about the sanction for missed appointments, he can start to see more context about how the app works. Here he can see that Mary’s missed appointments come after an update to the API used in the machine learning algorithm. He has the option to flag this data, escalating it within the branch.
4.7 Talking to Mary, she explains that the service has recently started suggesting appointments that she is having difficulty getting to on time. While the suggested days of the week are acceptable to her, the suggested time slots no longer fit with her journey from her children’s school. This hasn’t been a problem until recently.

4.8 Subhit notices the correlation with the change in the machine learning algorithm and flags Mary’s case for investigation. This is only possible because the data is structured in a way that allows this information to be shared and examined.

5 Scenario 3: Investigating when things go wrong

5.1 How parliament and other organisations investigate and audit digital services that rely on automated decision-making is an important question. They will need access to specialist skills to help them understand how algorithms are functioning. To enable this, digital services will need to make available new types of information for parliament to have a complete picture of how a service works over time.

5.2 A parliamentary committee is reviewing the effectiveness of the National Benefit Service.

5.3 Citizens Advice and other organisations have reported a spike in sanctions for missed appointments in Greater Manchester and the issue has been covered in the local press. The parliamentary committee wants to understand what might have gone wrong.
5.4 The committee has a data scientist on staff, Abbie, who can help members of the committee to understand how the service works and run simulations.

5.5 Abbie has access to lots of information and resources that the National Benefit Service makes available online to aid investigations such as this. These include:

- The release notes for each version of each algorithm that runs the service, and the ‘commit’ messages developers write to explain code changes
- A sandbox to explore different data inputs for each version of the algorithms
- Data about high-level trends over time
- The software tests that explain how the algorithms should work
- The test data used to train machine learning algorithms
- The source code of non-machine learning algorithms
- A log of when the algorithms were previously audited
- Screen captures of each version of the user interface
- Information about where data gets stored and processed

5.6 Abbie is able to test the correlation between the date of the change to the service and the spike in sanctions. Investigating further, Abbie can see that the changes made to the service include a change in the source of travel data used to train a new version of the machine learning algorithm. This change was made at a national level. They can see that this new API was tested, but only in London, Birmingham and Bristol. Similar spikes in sanctions now exist in cities where there is less real-time open transport data available.
5.7 By testing different versions of the algorithm, Abbie is able to conclude that it has affected the appointment times given to claimants.

5.8 It’s here that the interplay between algorithms affects claimants. The machine learning algorithm doesn’t adjust quickly enough to offer claimants like Mary suitable times for her appointments. Meanwhile, her missed appointments trigger another algorithm which tells someone at the claim office that she’s met the criteria for a sanction. No single algorithm is ‘at fault’, but claimants’ lives might be dramatically affected by the combination.

5.9 A committee examining this issue would have to understand where to intervene. Should they recommend wider testing of public services? Should they recommend a change to the guidance applying to sanctions when algorithms are updated? How should changes be communicated to claimants to mitigate the impact of changes?

5.10 One thing is clear: only with transparent, accountable systems can users of these services be adequately supported.

6 Designing to explain automated decisions

6.1 High-level principles for transparency, accountability, expertise and design will be critical for ensuring good practise in automated decisions. In concert, these will ensure that services which rely on algorithms can be implemented fairly, for public benefit.

6.2 First, services based on the outcome of automated decisions need to explain how these decisions have been reached. The frustration and disengagement people experience with these services stem from the fact that, all too often, services are opaque. Public services in particular can feel like an untouchable collection of decisions and rules which no one is allowed to understand. If we allow decisions made by machines to continue this trend, government risks further undermining the relationship between citizens and the state.

6.3 Second, services based on the outcome of an algorithm need to empower users to raise a complaint if something isn’t right. It’s easy to think of times an app has displayed the route to a shop, only to have picked a one-way street. Or a time a related link on a news article has picked up on the wrong keywords. These are relatively minor inconveniences. But what about when a service issues a traffic ticket to someone who shares the same name and birth date with the intended recipient? Without a way of flagging something that seems wrong, people will be at the mercy of decisions made by the services they use.

6.4 Parliament also needs to ensure that it has access to the skills required to understand how automated decisions come to be made. This will require a combination of new roles, new tools and new relationships with research institutions. It will mean understanding and specifying what information a service needs to log or publish, and what the appropriate mechanisms are for doing so. Not all algorithm inconsistencies will get the attention of a parliamentary committee, but they should be discoverable by empowered citizens.
6.5 Finally, we think context is critical. How a service explains its workings to users will be different to how it explains its workings to auditors. A suite of design patterns, data models and tools will be necessary to establish trust in automated decisions. As our example shows, service users, advice groups and parliamentary investigators have fundamentally different needs when it comes to understanding automated decisions.

6.6 The approaches we’ve outlined in our example scenarios are just a few of the available design solutions. We’ve chosen them because they show how diverse the approaches could be, not because we believe they’re the best solutions.

6.7 Instead, we advocate that services should be designed with the needs of all of these groups in mind, and that the algorithms that power such services should be open to interrogation by any one of them.

6.8 Only by thinking about these factors – transparency, accountability, expertise and design – will we see services powered by algorithms attain the level of trust needed to power public services.

April 2017