Cabinet Office

Version 1.0
Published 19 May 2016

# Data Science Ethical Framework

Data science carries both huge opportunities and a duty of care. Technology is changing so rapidly; as are the public's views. In this new and changing landscape, this document is not about creating additional hurdles, but rather about making innovation easier. It does this by bringing together the relevant law in the context of new technology, and prompting consideration of public reaction so that government data scientists and policymakers can be confident to innovate appropriately with data.

Developing the ethics around data science can't be done by government alone. This framework is a first iteration - a beta, if you like - of a set of principles wider than the legal framework, to help stimulate innovative and responsible action.

I look forward to listening to, and participating in that debate.

**The Rt Hon Matt Hancock MP**
**Minister for the Cabinet Office and**
**Paymaster General**

# Contents

# Why data science ethics are important

### Who is this guidance for?

This guidance gives those analysing or making policy or operational decisions with data the confidence to innovate. It balances the use of new data and techniques with respect for privacy and makes sure no-one suffers *unintended* negative consequences. An introduction to data science can be found here.

### Why is guidance needed for data science?

Data science is a new practice for government which provides opportunities to create insight and improve public services. Digital advances are producing huge amounts of new forms of data, allowing computers to more quickly process this data and makes decisions without human oversight. This creates new opportunities and many new challenges we have not had to consider before.

The law (e.g. the Data Protection and Intellectual Property Acts) sets out some important principles about how you can use data. And analytical, health and other professions have high standards for the quality and integrity of data processes. Those working with data should be aware of these and always act within them. But these are often in different places and not written with data science in mind. This guidance gives people the confidence to innovate by bringing together these laws and standards in the context of the rapidly evolving data landscape.

Public attitudes to data are changing. Working with data in a way which makes the public feel uneasy, without adequate transparency or engagement, could put your project at risk and also jeopardise other projects across government. Consideration of public attitudes and communication with them is key: most people are data pragmatists if told how society will benefit and how risks are managed.

Rather than creating additional hurdles this guidance makes it easier to innovate by helping you both navigate the legal aspects applicable to data science and think through some of the ethical issues which sit outside the law.

### How to use the guidance

Data science projects have a number of stages; discovery work to explore what it is possible to do with the data; the actual delivery; refining the accuracy of the insight; and the ongoing use of that insight by policymakers or operational staff. This guidance will help you think through the methodology and ask appropriate questions about how the project is conducted at each stage.

The guidance gives six principles which are based on existing law. **Fundamentally, the public benefit of doing the project needs to be balanced against the risks of doing so.**

1 Start with clear user need and public benefit

2 Use data and tools which have the minimum intrusion necessary

3 Create robust data science models

4 Be alert to public perceptions

5 Be as open and accountable as possible

6 Keep data secure

The guidance starts with a summary and checklist against the six principles and then goes on to explain each principle in more detail with real examples of where data science has been used well and less well, and practical suggestions of what you can do to act ethically. The Information Commissioner's Office has confirmed that the checklist can form the basis of a Privacy Impact Assessment.

This guidance is based on existing law. The exemptions within the Data Protection Act around crime, fraud and national security still apply.

The guidance will be iterated and developed with feedback from departments and external stakeholders. It is designed to be iterated as it is used, and is shared in the expectation that it will encourage feedback and further improvement. It also complements other ethical frameworks for analysis such as those relating to health data and from the National Statistician.

# Six key principles: at a glance view

## 1 Start with clear user need and public benefit

Data science offers huge opportunities to create evidence for policymaking, and make quicker and more accurate operational decisions. Being clear about the public benefit will help you justify the sensitivity of the data (principle 2) and the method that you want to use (principle 3).

## 2 Use data and tools which have the minimum intrusion necessary

You should always use the minimum data necessary to achieve the public benefit. Sometimes you will need to use sensitive personal data. There are steps that you can take to safeguard people's privacy e.g. de-identifying or aggregating data to higher levels, querying against datasets or using synthetic data.

## 3 Create robust data science models

Good machine learning models can analyse far larger amounts of data far more quickly and accurately than traditional methods. Think through the quality and representativeness of the data, flag if algorithms are using protected characteristics (e.g. ethnicity) to make decisions, and think through unintended consequences. Complex decisions may well need the wider knowledge of policy or operational experts.

## 4 Be alert to public perceptions

The Data Protection Act requires you to have an understanding of how people would reasonably expect their personal data to be used. You need to be aware of shifting public perceptions. Social media data, commercial data and data scraped from the web allow us to understand more about the world, but come with different terms and conditions and levels of consent.

## 5 Be as open and accountable as possible

Being open allows us to talk about the public benefit of data science. Be as open as you can about the tools, data and algorithms (unless doing so would jeopardise the aim, e.g. fraud). Provide explanations in plain English and give people recourse to decisions which they think are incorrectly made. Make sure your project has oversight and accountability built in throughout.

## 6 Keep data secure

We know that the public are justifiably concerned about their data being lost or stolen. Government has a statutory duty to protect the public's data and as such it is vital that appropriate security measures are in place.

More detail in annex below

# Quick checklist

← - - - - Tick where you are on the scale - - - - - →

| | | | | | |
|---|---|---|---|---|---|
| **1. Start with clear user need and public benefit** | **A.** How does the department and public benefit? | ○ ○ ○ ○ ○ ○ | High public benefit (to society or to an individual) | Medium public benefit (to society or to an individual) | Low public benefit (to society or to an individual) |
| **2. Use data and tools which have the minimum intrusion necessary** | **B.** How intrusive and identifiable is the data you are working with? | ○ ○ ○ ○ ○ ○ | Non-personal and therefore non-identifiable | Personal but non-sensitive | Personal, sensitive data which could be inferred or directly re-identified |
| | **C.** If identifying individuals, how widely are you searching personal data? | ○ ○ ○ ○ ○ ○ | Querying against known individuals | Querying against a targeted group | Speculatively searching for needle in haystack |
| **3. Create robust data science models** | **D.** What is the quality of the data? | ○ ○ ○ ○ ○ ○ | Representative and unbiased | Historical data which is biased and excludes certain groups | Inaccurate or missing data |
| | **E.** How automated are the decisions? | ○ ○ ○ ○ ○ ○ | Human making decision based on analysis | Limited human oversight but regularly checked | No human oversight or method of checking |
| | **F.** What is the risk that someone will suffer a negative unintended consequence as a result of the project? | ○ ○ ○ ○ ○ ○ | Low | Medium | High |
| **4. Be alert to public perceptions** | **G.** If personal data for operational purposes, how compatible was it with the reason collected? | ○ ○ ○ ○ ○ ○ | Very compatible | Less compatible but fair | Not compatible |
| | **H.** Do the public agree with what you are doing? | ○ ○ ○ ○ ○ ○ | Yes | Some would, some wouldn't or not sure what people think | No, or lots would have real concerns |
| **5. Be as open and accountable as possible** | **I.** How open can you be about the project? | ○ ○ ○ ○ ○ ○ | Very open, and make open the tools and data for re-use | Open about project but not about data/tools | Cannot talk about project aim |
| | **J.** How much oversight and accountability is there throughout the project? | ○ ○ ○ ○ ○ ○ | Throughout - including the decision made as a result of insight | Only at the beginning | None |
| **6. Keep data secure** | **K.** How secure is your data? | ○ ○ ○ ○ ○ ○ | Very secure, with restricted access to a few named individuals | Secure and password protected | Openly available within the department |

Some departments might find themselves at the left hand side of the scale, and others more on the right (blue), reflecting the nature of their department's work. This does not mean the project should not go ahead, but think carefully about it, and if possible, bring some elements to the green end of the scale.

*Not all may apply to your project

| | | |
|---|---|---|
| **All fine?**<br>**Go forward!** | **Some issues?**<br>**Think carefully** | **Tricky issues?**<br>**Extreme care & oversight** |

## Tricky issues?

Extreme care & oversight

## Answering these questions will also act as your Privacy Impact Assessment

**1. Start with clear user need and public benefit**

How does the public benefit outweigh the risks to privacy and the risk that someone will suffer an unintended negative consequence? (PIA Step 1)

Brief description of the project, including data to be used, how it will be collected and deleted. (PIA Step 2)

What steps are you taking to maximise the benefit of the project outcome?

**2. Use data and tools which have the minimal intrusion necessary**

What steps are you taking to minimise risks to privacy? (for example using less intrusive data, aggregating data etc)?

**3. Create robust data science models**

What steps have you taken to make sure the insight is as accurate as possible and there are minimal unintended consequences? (for example thinking through quality of the data, human oversight, giving people recourse)

**4. Be alert to public perceptions**

How have you assessed what the public or stakeholders would think of the acceptability of the project? What have you done in addition to the above to address any concerns?

Risks (PIA Step 3) and mitigating steps (PIA Step 4)

**5. Be as open and accountable as possible**

How are you telling people about the project and how you are managing the risks?

Who has signed this off within your organisation? Who will make sure the steps are taken and how? PIA Step 5

**6. Keep data secure**

What steps are you taking to keep the data secure?

# Annex

## each principle in detail, plus real life examples

# 1 Start with clear user need and public benefit

## Background

Data science offers huge public benefits in creating better evidence-based policy, making government operations more targeted and efficient, and keeping people safe.

Some data science projects will have a direct and tangible benefit to individuals, and some will improve policymaker's understanding so they can develop better policy. Creating a use case means that you can translate why better understanding will have benefits for individuals.

Understanding public benefit and creating a use case will allow you to:

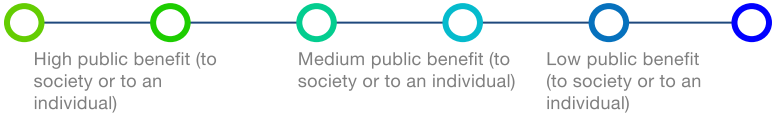
- Consider what risks it justifies and therefore what data and method you should use;
  - The risk to privacy **(Principle 2)**
  - The risk of making mistakes and negative unintended consequences **(Principle 3)**
- Start to think about what decisions might be taken as a result of the insight.

The public cannot easily distinguish between the ethics of data science (the production of the insight) and the decision or intervention taken as a result. They are more likely to be content if it is a supportive intervention rather than a punitive one (unless someone has broken the law). Therefore you need to consider the way the public benefit will be achieved as well as the data science method.

You need to have an idea of the probability of achieving the public benefit and identify metrics for assessing this at the start.

## Questions

Checklist question A: How does the department and public benefit?


High public benefit (to society or to an individual)
Medium public benefit (to society or to an individual)
Low public benefit (to society or to an individual)

*Further questions:*

- *Policymaking: If it is to differentiate or target a policy, what form might that take (supportive or punitive)?*
- *Operational: If it is to identify an individual, what level of harm could they cause or what would be the benefit of their identification?*

**Case studies**

**Good** – The ONS project on mobile phone data will help government understand traffic congestion and allow better road planning. For an individual user, this means they will get to work on time.

**Caution!** – A company analyses the data of customers who call into a call-centre and uses data from commercial databases, social media catalogues and census archives to match them with an agent with a similar personality. It is not clear whether the caller gives consent for this and it seems unlikely that the caller is aware of all the data being used, the assumptions being made or the storage of their data. Whilst this is openly available data, this may not pass the acceptable level of intrusion for most people for the service they are getting in return.

# 2 Use data and tools which have the minimum intrusion necessary

### Background

We can use more data, from more sources, more quickly than ever before - data science can help us collect this data, but also use it to its full advantage.

Existing law (the Data Protection Act) sets out how you can use personal data, but applies a minimisation principle, i.e. only use the minimum data you need to achieve the project aim. Ways to do this include:

- Only use personal data if similar insight or statistical benefit cannot be achieved using non-personal data
- De-identify individuals or aggregate to higher geographical levels where possible
- Query against datasets through APIs rather than having access to the whole data set
- Use synthetic data to get results

In order to build accurate data models, the data needs to be as representative and accurate as possible, so sometimes you have to include everyone's data. Ideally personal data would be de-identified, but in some instances (for example identifying terrorists or criminals) you may want to go on to identify individuals. In these cases, it would be best to train the algorithm on a small set of data about people of interest (e.g. criminals) and then apply this to larger datasets so that we do not miss anyone.

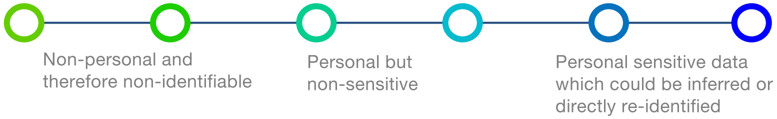
Using data that is voluntarily in the public domain (e.g. social media data) needs careful consideration. Legally it is personal data and needs to be processed fairly (i.e. in line with the T&Cs of the social media provider). It might be more appropriate to use social media data to spot trends or clusters of activity and alert local service providers to take action, than to take action yourself, but this would depend on the level of public benefit, level of consent and context in which it was provided (e.g. expectations of how a tweet is used is probably very different from sensitive discussion on Mumsnet - although both are publicly available).

**Data also has to be legally collected, stored, shared processed and deleted. The Data Protection Act sets out guidance on this.**

The law states that you must take reasonable steps to ensure that individuals will not be identifiable when you link data or combine it with other data in the public domain. The increasing number of datasets available now or in the future means that it might be easier to link to other open data sources to infer an individual's identity or personal information about them.

### Questions

Checklist question B: How intrusive and identifiable is the data you are working with?

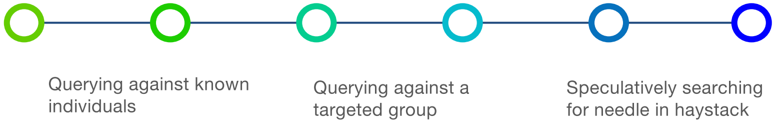


Non-personal and therefore non-identifiable

Personal but non-sensitive

Personal sensitive data which could be inferred or directly re-identified

Checklist question C: If identifying individuals, how widely are you searching personal data?



Querying against known individuals

Querying against a targeted group

Speculatively searching for needle in haystack

*Further questions:*

- *How can I meet the project aim using the minimum intrusion possible?*
- *How (re-)identifiable is the data?*
- *How much would people care about the data (how sensitive is it)?*
- *If using data that the public have freely volunteered, would your project jeopardise people providing this in the future?*

Information Commissioner's Office definitions about personal and personal sensitive data

# 2 Use data and tools which have the minimum intrusion necessary

**Case studies**

**Good** – DWP and DECC were using energy efficiency data and winter fuel allowance data to identify areas which were fuel poor but energy inefficient so they could better target efficiency advice. They had data which allowed them to identify individual households, but considered this too intrusive, so aggregated it up to post-code level.

**Good** – Streetbump is a mobile crowdsourcing app that records motion in cars when going over a pothole. This data is then sent back to the council to fix the pothole in that area. The data is collected by anyone with a smartphone who uses the app in their car. However, this could potentially cause certain areas of the city where more people use smartphones, to have their roads repaired whilst those with no access to smartphones are stuck with the potholes. Streetbump noticed this issue before the app launched and so first deployed it to city road inspectors, who service all parts of the city equally and asked the public to provide additional supporting data.

**Caution!** – A charity developed a tool that analyses the sentiment of a user's tweets, assesses whether they are suicidal and alerts friends who have signed up to use the tool so that they can talk to them. Whilst this is an admirable project and potentially life-saving tool, there was a backlash on social media which resulted in the app being taken down. Although this data is openly available for anyone to see, the idea of it being used for such a purpose proved ultimately too controversial.

**Web scraping guidance**

- Always respect website terms and conditions & robots protocol
- Notify website owners of any plans to scrape their websites on a large scale
- Schedule web scraping activities so as to minimise the impact on target websites
- Do not scrape website anonymously - make sure an identifiable IP address is visible
- Obtain explicit agreement from the website owner for scraping a website for statistical production purposes
- Ensure that any republishing data sourced from the web could not be interpreted as a breach of intellectual property rights

# 3 Create robust data science models

**Background**
Machine learning techniques can analyse far larger amounts of new types of data far more quickly and far more accurately than humans. Machine learning algorithms have the potential to vastly improve operational and policy decisions and provide a more personalised service to citizens.

But these tools are dependant on the input data and do have limits. Existing law (the Data Protection Act) states that you should not use personal data in a way that *unjustifiably* adversely affects someone. These questions will help you work out the risk to your project of unintended consequences.

Limits and opportunities of data

Algorithms learn from large amounts of historical data to make decisions in a way that traditional techniques cannot. However, the quality of this historical data can affect algorithms and can reinforce bias. If there are errors or bias in your input data, the algorithm will just perpetuate this unless it is recognised in advance. Use techniques such as shadow analysis to spot bias and then code in affirmative action to remove bias.

Equally, some new forms of data may give valuable real-time insight, but not be representative of the whole population. For example digital media is not representative of the entire population; there are some people without smartphones or who do not want to make their views known on the internet. This doesn't mean you shouldn't use this data, but that you should include metrics about who the data is representative of.

Algorithms can search through a huge amount of variables and identify which ones are most important in making certain decisions. We might not always know which ones they pick, (which is different from regression analysis where we do). Even if we remove prohibited variables (e.g. ethnicity), others might act as proxies for them. E.g. an algorithm which excludes people from certain geographical locations might also end up excluding ethnic minorities who live there.

Limits and opportunities of tools

Algorithms can prove to be much more accurate than human decision makers in many use cases, but they are not 100% accurate.

Data science can gain new insight from existing data, rather than collecting new data to answer each new question. Some of this data will be inferred by analysing single data sets (e.g. searching on Google for flu symptoms is a good indicator of having the flu), or linking data. Sometimes sensitive personal data can be inferred from personal data (for example health conditions from shopping habits). The Data Protection Act states that any additional personal data that is created needs to be accurate, so creating robust models is even more important in these cases.

Machines make trade-offs to find their target answer; either by narrowing down criteria or groups, but not identifying all data points that meet that criteria (false negatives) or widening out criteria or groups, but identifying some incorrect data points (false positives). The public are concerned about missing false negatives where this affects public safety, and also about including false positives when there is a negative punitive intervention for that person.

It is difficult to translate a complicated policy into a coded algorithm to make operational decisions (for example benefit applications), and new data will change the way some algorithms make decisions to get to their target answer.

Quality assurance

Be honest about the results of the data science, particularly if you're using data sources and techniques mentioned above. Create an accuracy rating for the insight so that people can decide how to use it (e.g. as an insight for a human decision, or a fully automated one). Make sure you understand the provenance of any derived insight that has come from elsewhere, and make sure this – and the accuracy rating - stays with it as it gets passed on to policy or operational colleagues.

# 3 Create robust data science models

Data scientists, policy or operational professionals should work together to design a question which is free from bias, check that the algorithm is correctly designed (so it accurately reflects the policy) and regularly test it especially when new data is added. Models that change and adapt are always at risk of breaking. Make sure you have procedures in place so you can tell when something has gone wrong and there is a process for fixing it.

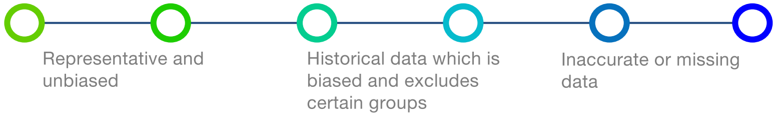
Unintended consequences

It is important to think through whether an incorrect decision will have unintended negative consequences for someone or distress someone, and if there is a risk - whether the public benefit justifies this.

In summary, different machine learning techniques work best with different projects. Automated machine learning works best where the policy can be easily translated into an algorithm, where you have high confidence in the data and where there is clear recourse for a decision made about an individual. In other cases, there should be more human oversight and the result should be used with other insight to inform the decision.

**Questions**

Checklist question D: What is the quality of the data?

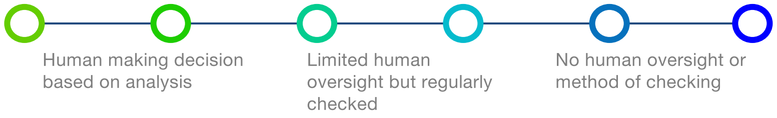


Representative and unbiased

Historical data which is biased and excludes certain groups

Inaccurate or missing data

Checklist question E: How automated are the decisions?

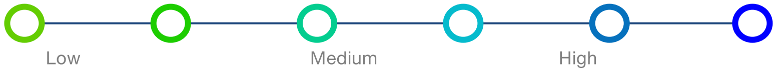


Human making decision based on analysis

Limited human oversight but regularly checked

No human oversight or method of checking

Checklist question F: What is the risk that someone will suffer a negative unintended consequence as a result of the project?



Low

Medium

High

*Further questions:*

- *Have you designed the algorithm with the policy team? How are the relevance criteria formed and how is data included or excluded? How might this lead to inherent bias within the system?*
- *How can you (or external scrutiny) check that the algorithm is still achieving the right output decision, especially when new data is added?*
- *What are the consequences of relying on the computer-made decision alone, and in what circumstances would additional human oversight be useful?*

# 3 Create robust data science models

**Good** – At CERN, two teams worked in isolation on the CMS and ATLAS detectors to search for the particle. Each team worked independently using different methodologies to ensure an accurate result and corroborate each other's findings

**Good** - As far back as 2002 a data science application was matched against a group of law professors and experts to predict the outcomes of the upcoming cases of the Supreme Court of the United States – the application scored 75%; the expert panel 59%.

**Good** - IBM's WatsonPaths has been developed to assist doctors make clinical decisions - the product has been developed to allow doctors to see the detailed reasoning to understand how the decision was made.

**Caution!** – A US state has a programme that makes automated decisions based on the data it collects. In one case, medical benefits were terminated for breast cancer patients based on income and asset requirements not required by federal or state law.

**Caution!** – A taxi firm has an automatic pricing algorithm which responds to surges in demand. During the hostage crisis in Sydney in December 2014, the algorithm raised the prices by up to four times the normal rate in the crisis area. They responded by saying that the algorithm was automatic and quickly brought the prices back to normal levels and recompensed their customers.

**Caution!** – A foreign country has a government service which identifies parents who owe money in child maintenance. The data matching process is often incorrect due to misspelled names or missing data which results in some individuals being incorrectly targeted automatically by the system with the result being a large bill, poor credit ratings and even freezing wages. The recourse for individuals who are incorrectly targeted is time-consuming and not straightforward.

**Caution!** – Analysis from social media surrounding Hurricane Sandy created the illusion that the centre of the disaster was in Manhattan. But this masked the really severely damaged locations of Coney Island, Breezy Point and Rockaway where few people had smartphones to contribute to the noise.

# 4 Be alert to public perceptions

**Background**
The law tells us what we can do, but ethics tells us what we should do. Ethics become more important when advances in technology are pushing our understanding of the law to its limits.

Ethics are people's moral understanding of what is right. Some of this is codified by law (e.g. anti-discriminatory practice) and some of this is not (e.g. smacking a child).

Both the law and ethical practice require us to understand public opinion so we can work out what we should do. The law itself (the Data Protection Act) also requires us to have an understanding of how people would reasonably expect their personal data to be used, particularly if for a different purpose. This can be a challenge as public opinion is diverse and is shifting over time.

The existing law does allow you to use personal data for research purposes even if it is different from the reason it was collected. Government has set up a number of 'safe havens' which allow government analysts and academic researchers access to such data.

Use of personal data for non-research purposes has to be fair, proportionate and compatible with the original purpose for which it was collected.

Understand both stated and revealed public opinion (people's actual behaviour) about how people would want the data you hold about them to be used. There are a number of different ways of understanding public opinion depending on the level of controversy of the project (see Open Policy Making Toolkit).

Make sure it is not just you making the decision and that you consult others to work out whether projects are acceptable**.**

**Different personas were derived from responses to a public dialogue on data science ethics and may help you to understand public attitudes to data science.**

**Questions**
Checklist question G: If personal data for operational purposes, how compatible was it with the reason collected?

○───○───○───○───○───○

Very compatible  Less compatible but fair  Not compatible

Checklist question H: Do the public agree with what you are doing?

○───○───○───○───○───○

Yes  Some would, some wouldn't or not sure what people think  No, or lots would have real concerns

*Further questions:*

- *How 'informed' was the consent given (to other data re-uses)?*
- *What research has been done into what people think about the problem you are trying to tackle and the data you want to use?*

**Case studies**
**Good** – when accessing data from third parties, you should be sure to explore all the legal issues. For example, a webscraping tool can allow you to access the data you need from a site, but if this is done without checking the sites exclusion protocols or T&Cs, you could access data or use it in a way that is not legal. For example, LinkedIn seems like a great resource to find data relating to jobs and careers of user, however LinkedIn have certain terms and conditions which prohibit scraping data from their user's profiles.

**Caution!** – A hospital used historic data on medical school applications decision to sort new applicants. The previous decisions used to train the model had unfairly discriminated against women and minority groups and so this prejudice was reflected in the new automated system.

# 5 Be as open and accountable as possible without putting your project at risk

**Background**

Transparency is essential to make the case for the benefits of data science and to avoid accusation of nefarious ‘secret’ big data projects. It is also a good antiseptic for unethical behaviour.

There will of course be some instances where we cannot be open about the work as doing so would jeopardise the project goal, e.g. allow people to understand how we are identifying illegal activity and game the system. The public understand these instances in principle, but they do need to be carefully considered.

Let people know about the social benefit of your work and the impact it has had on collective or individual social or financial outcomes.

Aim to be publicly transparent about what you are doing and be open about the tools, data and algorithms used and its intention (unless there are serious public interest reasons not to, such as fraud or counter-terrorism) and provide your explanations in plain English.

Try and ensure that for new digital services and elsewhere, where possible people can view and extract their own personal data held by the government service and that they are told about how their data will be used at the point of collection.

Make sure there is oversight and accountability throughout the project. Oversight needs to cover both the initial assessment of purpose and method, but then how it is carried out and what is done as a result. Oversight might vary depending on the level of controversy of the project. Projects that have no ethical issues may be overseen by the policy lead; projects that have several ethical issues may require external oversight.

Once you have considered the unintended consequences of a project (see principle 3), you should put in place a means of recourse through which people can challenge incorrect decisions. Another way of mitigating the risk of unintended consequences could be to give people choice in the decision made, for instance offering them several options for a personalised service, not just one.

**Questions**

Checklist question I: How open can you be about the project?

Very open, and make open the tools and data for re-use

Open about project but not about data/tools

Cannot talk about project aim

Checklist question J: How much oversight and accountability is there throughout the project?

Throughout - including the decision made as a result of insight

Only at the beginning

None

*Further questions:*

- *How are you explaining how you made your decision (or how the computer did) in plain English?*
- *How do the public know that data held about them is being used for a data science project, and to what extent can they see this data and correct if necessary?*
- *What is the process for overseeing the project?*

# 5 Be as open and accountable as possible without putting us at risk

**Case studies**

**Good** – The 100K Genome Project does have some ethical concerns about the use of the data – they cannot guarantee that the data they publish (although anonymised) could not be re-identified in the future using techniques not yet invented. However, the organisation are careful to be very clear about this information upfront so that the individual knows exactly what their data is being used or could be used for.

**Good** – Government collected the National Food Survey from 1940 to 2000. Defra wanted to publish it as open data so that other people could extract maximum possible value from it, for example school children keeping food diaries and comparing it to what their predecessors ate. But they had to do so while preserving confidentiality, particularly if it could be combined with other data in the public domain. So – bound by the Official Code of Statistics which has penalties for those who break the rules – they created some robust internal processes and published a Privacy Impact Assessment for others to comment on.

**Caution!** – A social media company have recently performed experiments on users by filling their newsfeeds with positive or negative stories and measuring the emotional reaction. It was not open with its users that it was doing this.

# 6 Keep data secure

**Background**

We know that the public are justifiably concerned about their data being lost or stolen and government departments have a responsibility to protect both personal and non-personal classified data. It is vital that we keep it secure. The Data Protection Act and government's own Security Policy Framework provide the basis of how data should be collected, stored, shared, processed and deleted. Departments will have their own guidelines for how data is shared and handled.

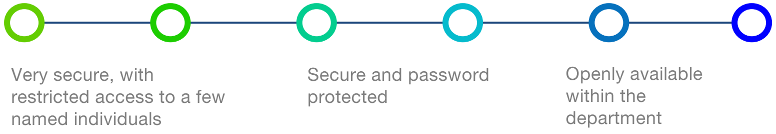
The Administrative Data Research Network has created a number of 'safe havens' where administrative data (data routinely collected by government) can be anonymised and linked, with strict controls for who has access for the data and for how long. Government researchers can access this resource and research has shown that the public supports this type of controlled analysis.

The Government Digital Service are creating canonical registers of data. Registers and APIs allow data to be held separately (distributed) and drawn together rather than in bulk data sets, making the data safer from attack.

The Data Protection Act outlines best practice for the retention and deletion of data. The ICO also provide guidance on how to best delete records.

**Questions**

Checklist question K: How secure is your data?



Very secure, with restricted access to a few named individuals

Secure and password protected

Openly available within the department

*Further questions:*

- *Who has access to the data and how? Are they aware of ethical considerations and frameworks?*
- *For how long is the data stored?*