Scenario 1 - stroke

Scenario

There are more than 100,000 strokes in the UK each year – that is around one stroke every five minutes. About 11% of patients die immediately or within a few weeks as a result of the stroke, making stroke the fourth biggest killer in the UK. Almost two thirds of stroke survivors leave hospital with a disability.

Rapid and accurate diagnosis of stroke greatly increases chances of survival and recovery of the patient. This is highly specialised work which ideally should be done by neuroradiologists with many years of training and experience. However, these experts are not available in each hospital, 24 hours a day, 7 days a week, and in practice diagnosis is often done by non-specialist emergency medicine doctors.

As diagnostic data are accumulated from previous stroke patients, automated decisions systems could provide stroke diagnosis that is fast, and always available in each hospital.

There are three automated decision systems for the NHS to choose from – system A, system B, and system C. Each system uses information about a patient's acute symptoms (for instance paralysis and loss of speech), their medical history, and neuroradiological images (such as CT-scans of the brain) to identify patterns that indicate whether he or she has had a stroke; the type of stroke; its location; and its severity.

- System A – Expert System

  This system uses an algorithm that was developed with help from experienced neurologists and neuroradiologists, and aims to follow the same reasoning as they would do. In practice it does not reach the same level of accuracy as they would, but the algorithm is completely transparent in the way it reaches its conclusions: for each individual case it can provide specific rules that were applied to reach a conclusion. It has an overall accuracy rate of 75%, which is comparable to what most emergency medicine doctors would achieve.

  This means that in 25% of cases, someone might be classified as having a stroke while they were not or vice versa, or the type, location, and severity of the stroke might be misjudged.

- System B – Conventional machine learning

  This system uses an algorithm that was established through machine learning from a large set of patient data, collected at English hospitals. This algorithm reaches (human) expert level performance, but it is not very

transparent in the way it reaches its conclusions: it can only tell us which features, in general, are important and which are not.

It has an overall accuracy rate of 85%. This means that in 15% of cases, someone might be classified as having a stroke while they were not or vice versa, or the type, location, and severity of the stroke might be misjudged.

- System C – Deep Learning

This system uses advanced AI derived from the same set of patient data as System B. However it has "taught itself" from the data which features were best able to distinguish strokes from non-strokes, and best able to distinguish different types of stroke, their location, and their severity. This algorithm is not transparent in the way it reaches conclusions: it is unable to provide any explanation that is understandable by humans.

However it has an overall accuracy rate of 95%, which is better than human experts perform. This means that in 5% of cases, someone might be classified as having a stroke while they were not or vice versa, or the type, location, and severity of the stroke might be misjudged.

	System A	**System B**	**System C**
Accuracy	75% (A&E doctor's level)	85% (human expert level)	95% (beyond human level)
Transparency	Full explanation	Partial explanation	No explanation

Scenario 1: Questions

1. How important is it for a patient to receive an explanation of an automated decision about stroke diagnosis?

   ☐ Very important
   ☐ Fairly important
   ☐ Not very important
   ☐ Not at all important
   ☐ Don't know

2. If system C was chosen by the NHS, almost no explanation would be provided. How much does this matter?

   ☐ Very much
   ☐ Quite a lot
   ☐ Not very much
   ☐ Not at all
   ☐ Don't know

   Why (up to three reasons)?

3. Which automated decision system do you think the NHS should choose? Explain the factors affecting your choice.

   System A – Expert System
   System B – Conventional machine learning
   System C – Deep Learning

Scenario 2: Recruitment

Scenario

When running recruitment campaigns, a private sector organisation currently tasks its staff with manually screening all job applications received, and making decisions about which applications to shortlist for interview. This often involves several members of staff reading through thousands of job applications. Sometimes, the organisation receives so many applications that its staff cannot (or do not) properly review every application.

So that the organisation is able to screen every application and free up staff to focus on other work, for future recruitment campaigns, it plans to use an automated decision system to screen job applications and make shortlisting decisions.

The automated decision system will use data about existing employees to be programmed, or to "learn", which qualities contribute towards being a high performing employee. This may include traditional qualities such as relevant experience, skills and qualifications. But there may also be other qualities that the system is programmed, or "learns", to treat as important such as particular writing styles, personality traits and personal interests.

The system will then be used to screen the CVs, covering letters and application forms of individuals applying for jobs. It will predict the likelihood of applicants becoming high performing employees. Based on these predictions, the system will decide whether to place them in 1 of 2 classifications:

- *Application accepted*

  Applicants predicted as likely to become high performing employees are placed in this classification and are accepted for interview.

- *Application rejected*

  Applicants predicted as unlikely to become high performing employees are placed in this classification and are rejected for interview.

There are 3 automated decision systems for the organisation to choose from:

- *System A – Expert System*

  This system uses an algorithm that was developed with help from experienced recruitment officers, and aims to follow the same reasoning as they would do. In practice it does not reach the same level of accuracy as they would, but the algorithm is completely transparent in the way it reaches its conclusions: for each individual case it can

provide specific rules that were applied to reach a conclusion.

When tested on existing data about recruitment, this system was shown to have an overall accuracy rate of 75%. This means that 25% of the time its predictions were incorrect (e.g. predicting that an applicant would be unlikely to become a high-performing employee when in reality they did, or vice versa).

The accuracy of this system is comparable to that of a typical recruitment officer.

- *System B – Conventional machine learning*

This system uses an algorithm that was established through machine learning from a large set of recruitment data, collected by the organisation. This algorithm achieves (human) expert level performance, but it is not very transparent in the way it reaches its conclusions: it can only tell us which features, in general, are important and which are not.

When tested on existing data about recruitment, this system was shown to have an overall accuracy rate of 85%. This means that 15% of the time its predictions were incorrect (e.g. predicting that an applicant would be unlikely to become a high-performing employee when in reality they did, or vice versa).

The accuracy of this system is comparable to that of a very experienced recruitment officer.

- *System C – Deep Learning*

This system uses advanced AI, derived from the same set of data as System B. However it has "taught itself" from the data. This algorithm is not transparent in the way it reaches conclusions: it is unable to provide any explanation that is understandable by humans.

However it has an overall accuracy rate of 95%, which is better than human experts perform. This means that 5% of the time its predictions were incorrect (e.g. predicting that an applicant would be unlikely to become a high performing employee when in reality they did, or vice versa).

	System A	**System B**	**System C**
Accuracy	75% (recruitment officer level)	85% (human expert level)	95% (beyond human level)
Transparency	Full explanation	Partial explanation	No explanation

Scenario 2 Questions

1. How important is it for an applicant to receive an explanation of an automated decision about accepting / rejecting a job application?

   ☐ Very important
   ☐ Fairly important
   ☐ Not very important
   ☐ Not at all important
   ☐ Don't know

2. If system C was chosen by the organisation, almost no explanation would be provided. How much does this matter?

   ☐ Very much
   ☐ Quite a lot
   ☐ Not very much
   ☐ Not at all
   ☐ Don't know

   Why (up to three reasons)?

3. Which automated decision system do you think the company should choose? Explain the factors affecting your choice.

   System A – Expert System
   System B – Conventional machine learning
   System C – Deep Learning

Scenario 3 - kidney transplantation

Scenario

Chronic kidney disease (CKD) is a condition characterized by a gradual loss of kidney function over time. It may be caused by diabetes, high blood pressure, and other disorders. If kidney function gets worse, wastes can build to high levels in your blood, leading to complications like anaemia (low blood count), weak bones, and nerve damage. About 2.6 million people (6.1%) in the UK live with CKD. There is no current cure for CKD.

Eventually, a person with CKD will develop permanent kidney failure, and they will need dialysis or a kidney transplant in order to survive. Dialysis is the removal of waste products and excessive fluids from blood using a machine, and typically needs to happen three times per week for at least 3 hours, placing an immense burden on the patient. A kidney transplant is a better option than dialysis, as patients will have a normally functioning kidney after the transplantation, enabling a relatively normal life.

Because there is a mismatch between demand for and supply of kidney transplants, patients often have to wait for many months (or even years) on dialysis before receiving a kidney transplant. There are currently about 8000 patients on this waiting list in the UK. There has to be a good 'match' between kidney donor and recipient in terms of blood type, immune system, and many other factors in order to maximise the chances of survival of the transplanted kidney. Determining whether donor and recipient match is usually done by experienced renal consultants. However in about 15% of cases the match is not ideal and the transplanted kidney does not survive more than 5 years – and some of them stop functioning within days or weeks. The transplanted kidney is removed and discarded and these patients will have to go back to dialysis.

The NHS wants to deploy an automated decision system for finding matches between kidney donors and recipients so as to make good use of the kidneys and avoid mismatches between donated kidneys and recipients. Each time a new donor becomes available, the system will use data about the kidney and about the potential recipient to determine, for each patient on the waiting list, whether the risk of transplanting the donor's kidney to that patient is 'high', 'intermediate', or 'low'. Only matches that are classified as low risk are eligible for actual transplantation. If the system indicates that there are multiple low risk matches for the same donor, younger patients will be prioritised.

It is hoped that with this system, a larger number of transplanted kidneys will survive longer. There are three automated decision systems to choose from – system A, system B, and system C.

- System A – Expert System

This system uses an algorithm that was developed with help from experienced kidney doctors, and aims to follow the same reasoning as they would do. In practice it does not reach the same level of accuracy as they would, but the algorithm is completely transparent in the way it reaches its conclusions: for each individual case it can provide specific rules that were applied to reach a conclusion.

It has an overall accuracy rate of 75%, which is a little lower than what is currently achieved in practice across the NHS (and lower than that achieved by the top specialists). This means that 25% of the time its predictions were incorrect (e.g. predicting that the kidney would last at least 5 years for the selected patient when in reality it didn't).

- System B – Conventional machine learning

This system uses an algorithm that was established through machine learning from a large set of patient data, collected at English hospitals.. This algorithm achieves (human) expert level performance, but it is not very transparent in the way it reaches its conclusions: it can only tell us which features, in general, are important and which are not.

It has an overall accuracy rate of 85%. This means that 15% of the time its predictions were incorrect (e.g. predicting that the kidney would last at least 5 years for the selected patient when in reality it didn't).

- System C – Deep Learning

This system uses advanced AI, derived from the same set of patient data as System B. However it has "taught itself" from the data which features were best able to distinguish successful matches from non-successful matches. This algorithm is not transparent in the way it reaches conclusions: it is unable to provide any explanation that is understandable by humans.

However it has an overall accuracy rate of 95%, which is better than human experts perform. This means that 5% of the time its predictions were incorrect (e.g. predicting that the kidney would last at least 5 years for the selected patient when in reality it didn't).

	System A	**System B**	**System C**
Accuracy	75% (below human expert)	85% (human expert level)	95% (beyond human level)
Transparency	Full explanation	Partial explanation	No explanation

Scenario 3 Questions

1. How important is it for a kidney patient and their family to receive an explanation of an automated decision about why the patient could or could not be matched?

   ☐ Very important
   ☐ Fairly important
   ☐ Not very important
   ☐ Not at all important
   ☐ Don't know

2. If system C was chosen by the NHS, almost no explanation would be provided. How much does this matter?

   ☐ Very much
   ☐ Quite a lot
   ☐ Not very much
   ☐ Not at all
   ☐ Don't know

   Why (up to three reasons)?

3. Which automated decision system do you think the NHS should choose? Explain the factors affecting your choice.

   System A – Expert System
   System B – Conventional machine learning
   System C – Deep Learning

Scenario 4: Criminal justice

Scenario

In an effort to reduce recurring low-level crime, a UK Police Force is setting up a rehabilitation programme to address the underlying issues that lead people to commit multiple minor offences (non-violent / non-sexual) and to prevent them from reoffending.

Artificial intelligence software will be used to identify the individuals to be offered a place on a rehabilitation programme rather than face trial. The individuals selected by the software will be those charged with a minor offence who are considered unlikely to go on to commit a serious offence in the next six months. Individuals that successfully complete the programme will not be prosecuted and will not receive a criminal conviction for the offence they were charged with, and no further action will be taken against them. However, if they fail to complete the programme, they are liable to face prosecution. Individuals offered a place on a rehabilitation programme may refuse it and choose to face prosecution.

The Police Force plans to use an automated decision system to make decisions about who to refer to the rehabilitation programme. It wants to include individuals likely to go on to commit further minor offences but exclude individuals likely to go on to commit a serious offence (violent / sexual).

The automated decision system will use information about an individual's offending history, age, gender and geographical area, as well as any available information about them from local agencies (such as social services) and national databases (such as the Police National Computer). It will analyse this information to predict the likelihood of the individual committing a serious offence over the next 6 months. Individuals will be placed in 1 of 2 classifications based on this prediction:

- *Eligible for rehabilitation programme*

  Individuals predicted as unlikely to commit a serious offence in the next 6 months are placed in this classification and referred to the rehabilitation programme.

- *Ineligible for rehabilitation programme*

  Individuals predicted as likely to commit a serious offence in the next 6 months are placed in this classification and referred for prosecution for the offence charged.

There are 3 automated decision systems for the Police to choose from:

- *System A – Expert System*

This system uses an algorithm that was developed with help from very experienced Police Custody Officers, and aims to follow the same reasoning as they would do. In practice it does not reach the same level of accuracy as they would, but the algorithm is completely transparent in the way it reaches its conclusions: for each individual case it can provide specific rules that were applied to reach a conclusion.

When tested on existing data about reoffending, this system was shown to have an overall accuracy rate of 75%. This means that 25% of the time its predictions were incorrect (e.g. predicting that an individual would commit a serious offence when in reality they didn't, or vice versa).

The accuracy of this system is comparable to that of an average Police Custody Officer.

- *System B – Conventional machine learning*

This system uses an algorithm that was established through machine learning from a large set of criminal offence data, collected by the police and local agencies. This algorithm achieves (human) expert level performance, but it is not very transparent in the way it reaches its conclusions: it can only tell us which features, in general, are important and which are not.

When tested on existing data about reoffending, this system was shown to have an overall accuracy rate of 85%. This means that 15% of the time its predictions were incorrect (e.g. predicting that an individual would commit a serious offence when in reality they didn't, or vice versa.)

The accuracy of this system is comparable to that of a very experienced Police Custody Officer.

- *System C – Deep Learning*

This system uses advanced AI, derived from the same set of data as System B. However it has "taught itself" from the data. This algorithm is not transparent in the way it reaches conclusions: it is unable to provide any explanation that is understandable by humans.

However it has an overall accuracy rate of 95%, which is better than human experts perform. This means that 5% of the time its predictions were incorrect (e.g. predicting that an individual would commit a serious offence when in reality they didn't, or vice versa.)

	System A	**System B**	**System C**
Accuracy	75 (Custody officer level)	85% (human expert level)	95% (beyond human level)
Transparency	Full explanation	Partial explanation	No explanation

Scenario 4 Questions

1. How important is it for an individual to receive an explanation of an automated decision about referral to a rehabilitation programme?

   ☐ Very important
   ☐ Fairly important
   ☐ Not very important
   ☐ Not at all important
   ☐ Don't know

2. If system C was chosen by the police force, almost no explanation would be provided. How much does this matter?

   ☐ Very much
   ☐ Quite a lot
   ☐ Not very much
   ☐ Not at all
   ☐ Don't know

   Why (up to three reasons)?

3. Which automated decision system do you think the police force should choose? Explain the factors affecting your choice.

   System A – Expert System
   System B – Conventional machine learning
   System C – Deep Learning