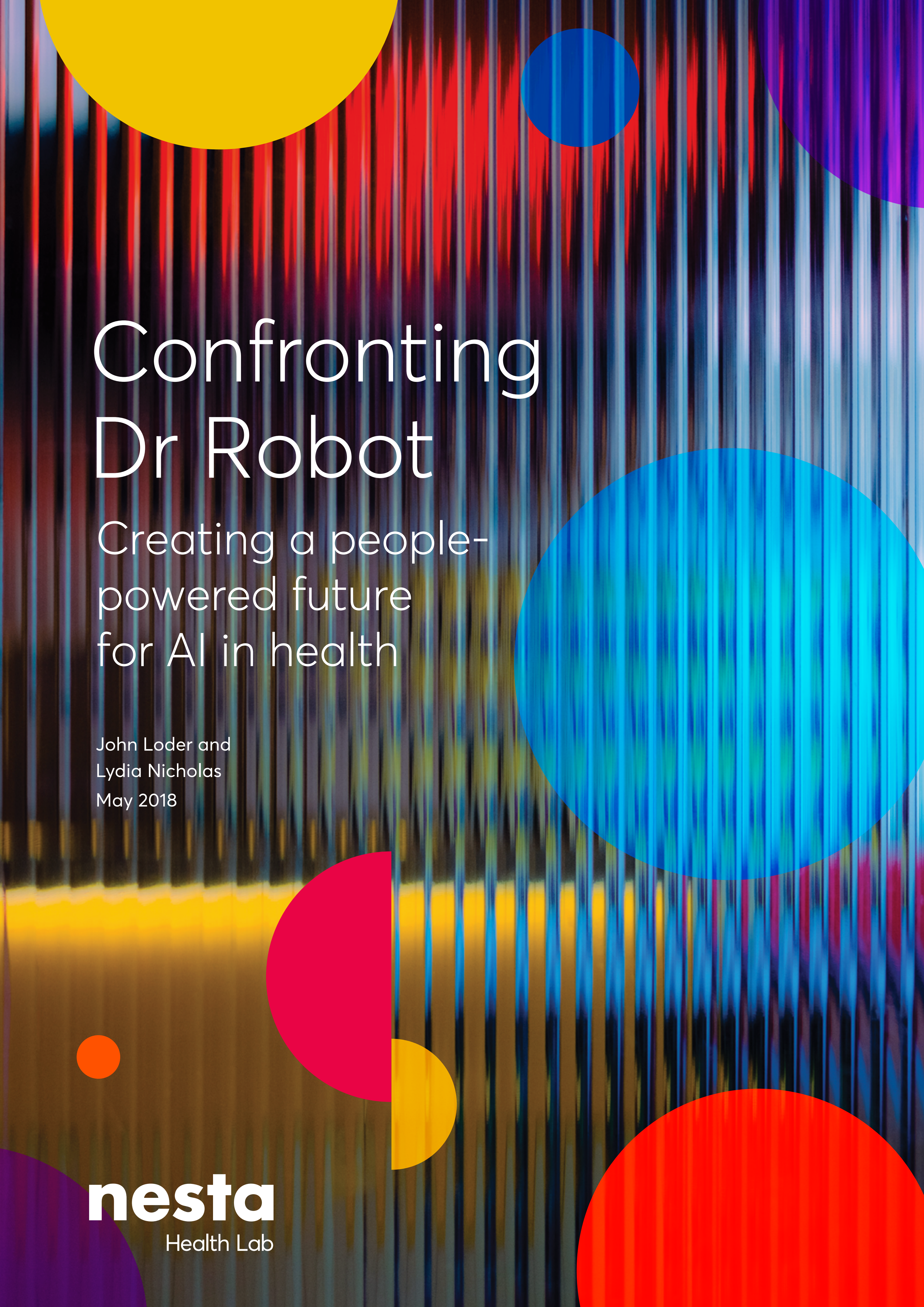

Confronting Dr Robot

Creating a people-powered future for AI in health

John Loder and
Lydia Nicholas

May 2018

nesta
Health Lab

Acknowledgements

We would like to thank the following for their insight and assistance with the report:

Geoff Huggins, **Dr Barbara Barry**, **Dr Harpreet Sood**, **Dr Thomas Hope**, **Dr Jeremy Wyatt**, **Will Cavendish**, **Saurabh Johri**, **Ben Glocker**, **Ben Hughes**, **Chris Tyson**, **Effy Vayena**, **Helen Munn**, **Indra Joshi**, **Alastair Pickering**, **Mark Hellaby**, **Tim Hubbard**, **Tom Dewar**, and **Zoe Camper**.

And our colleagues from Nesta, **Harry Armstrong**, **Halima Khan**, and **Celia Hannon**.

About Nesta

Nesta is a global innovation foundation. We back new ideas to tackle the big challenges of our time.

We use our knowledge, networks, funding and skills - working in partnership with others, including governments, businesses and charities. We are a UK charity but work all over the world, supported by a financial endowment.

To find out more visit **www.nesta.org.uk**

About Nesta Health Lab

Nesta Health Lab is committed to being a centre of expertise on people-powered and data-driven health: we work with partners from the health, care, voluntary, community and social enterprise sectors to test and scale new ways for people to remain healthy.

Our work focuses on new sources of support, which make it possible for people to be more involved in their health, such as peer support; new sources of data, which improve people's knowledge about their health, such as citizen-generated smartphone data; and new sources of innovation, that generate new solutions, including our 100 day People Powered Results method for transforming systems.

Over the past few years, we have backed over 100 local health systems and individual organisations with more than £20 million of funding.

www.nesta.org.uk/health-lab

nesta
Health Lab

Confronting Dr Robot

Creating a people-powered future for AI in health

May 2018

nesta Health Lab

Executive summary

Artificial Intelligence could become part of the front door to healthcare. It could make the health system simpler, more accessible, more responsive, more sustainable, and put patients more in control. But there's a risk that the public could experience it more as a barrier than an open door, blocking access to care, offering opaque advice and dehumanising healthcare in every sense. We're now at a crucial moment when decisions are being made which will determine whether the technology develops into People Powered AI.

Artificial intelligence (AI) looks like it could be one of the transformative technologies of our era. Healthcare is rich in the data that AI thrives on, and in the kinds of questions that it can tackle. While the use of AI in healthcare is at an earlier stage than the hyperbole surrounding the technology might suggest, it is developing at pace, and this raises both significant opportunities and risks.

AI has delivered some striking results. There have been research trials that successfully use machine learning on images from, for example, radiology, dermatology and ophthalmology, to a level of accuracy that matches clinicians' own abilities. This, and other AI developments, have led to the suggestion that machines are poised take the place of doctors.

However, today's AI is narrow and not capable of the holistic thinking and complex judgement required for many clinical tasks. While there are significant areas of medicine where more narrow applications of decision-making rules and expert pattern matching predominate, the path towards AI replacing humans is not solely determined by technical capability. Technology implementation will need to address trust, accountability and similar factors. And, at the same time, humans remain especially good at certain tasks, such as learning to identify rare situations from small amounts of data.

This jump to focusing on whether or not AI could replace doctors also potentially distracts from some far more immediate and likely applications of AI in health. It is far easier for AI to be adopted where there are no or few good alternatives on offer, than in areas where humans are effective and trusted. Areas where AI could be effective, and where there are few good alternatives include:

Advice and triage before seeing a doctor
When should someone first seek help from the health service?

Proactive care
When is the right time to intervene in the face of worsening symptoms?

Automated second opinion
How does the diagnosis and treatment I am getting compare to alternative options?

There are already AI products in the market in these areas, although the evidence base is not always sufficient. If AI is adopted in these areas it would be in a hugely influential position over our health and care. This could bring great benefits, but also comes with significant risks that need to be proactively managed and mitigated.

Applications of AI

1. Advice and triage before seeing a doctor

Most people find it hard to know exactly when to seek appropriate medical help. Twenty per cent of GP[1] appointments and 19 per cent of A&E attendances[2] are for minor medical problems that could be treated at home. This unnecessary or preventable demand creates significant pressure on health resources.

AI is already being used to offer healthcare advice and diagnoses directly to consumers. People can buy a diagnostic app or access to a chatbot to share symptoms with, and use this advice to decide whether or not to seek further medical help. In this way AI is beginning to provide a form of triage into the healthcare system and, if developed correctly, could help solve a major issue: how to support the appropriate use of limited healthcare resources including reducing the unnecessary use of health services.

However, this use of consumer-facing AI could also generate a flood of unnecessary demand (from false positives or generally risk averse advice); a new source of error in the system (from false negatives or other mistakes); and could widen health inequalities depending on the underlying business model. This area of development for AI also creates a new and urgent regulatory challenge.

These challenges need to be tackled, however, because this looks set to be a growth market and we are increasingly likely to see the use of AI as a highly influential entry point to the healthcare system.

A likely future is one where AI is a common first point of contact for health, and a front door to the health system - a highly influential position.

2. Proactive care

AI is also becoming capable of extracting signals from real-time data and giving early warning that a health problem is getting worse. For example, by listening to the breathing sounds of those with congestive heart failure to spot signs of deterioration. This could enable help to be directed to the people who need it in a more timely way, leading to a healthcare system that is more dynamic and responsive in the way it cares for people. Or it could generate a great deal of unnecessary concern, replace individualised conversations with standardised analytics, and generate an oppressive degree of monitoring.

AI would have an influence over who gets treated and when they get treated. While this could make the system much more dynamic, it could also make it more impenetrable, more unequal and less individualised.

3. Automated second opinion

AI could sit alongside doctors and offer an opinion on the same patient, for both diagnosis and treatment. This could be used to give patients a digital second opinion, which would enable to them to challenge and advocate for their care more easily. It would also allow mass comparisons of physicians' decisions centrally, perhaps with a view to understanding variation in care better. This use of AI could have significant potential influence on power dynamics within health, as well creating additional cost pressure on the system. The evidence that AI can do this reliably is not there, but products like IBM Watson claim to come close to this sort of functionality.

Both patients and managers will find it tempting to compare diagnosis and treatment to AI-generated opinions This could be helpful if applied to the narrower questions where AI is competent, but high risk if misapplied to judgements beyond the capability of the technology or quality of underlying datasets, putting pressure on clinicians to conform to it.

Consequences for power and autonomy

Like many digital technologies, AI can either democratise or centralise, empower or disempower, depending on the way it is implemented. It is easy to see how poorly designed and executed AI would be problematic:

- Poorly designed triage and prioritising systems could make the system even harder to access, making healthcare even **less simple** for patients.
- Increasing use of data-hungry analytics could **squeeze out dialogue**, which could make it harder to surface key details about the individual not captured in datasets.
- Opaque AI could **reduce accountability and transparency**.
- An over-monitored patient is not helped to understand what the AI is saying and why, making it harder to have a say in their care and leading to **reduced control**.

Further, it is likely that those worst affected by these changes will be those with hard to diagnose conditions, complex social and health needs, and who already face disadvantage.

However, it is equally possible to imagine a future where the patient is significantly more empowered:

- AI makes it **simpler** to know when to seek help and get to the right person.
- Ensure patient and professionals are more prepared for and have more time for their conversation, so **more dialogue**.
- Patients find it easier, via home diagnostics and chatbot advice to understand their condition and to ask for help when they need it, so have **increased control**.
- The ability to get a digital second opinion **increases transparency and accountability**.

Also the clinician could be freed from a lot of low priority work, fed useful insights, and more able to intervene at the right time.

Which of these futures turns out to be the real one is not purely determined by the technology itself, but by the choices that are made in its implementation. We need to be as careful in thinking about how new technology integrates with key relationships and pathways as with how it integrates key technical systems.

Therefore, in addition to core questions of whether it is **safe** and **effective**, we should also be applying the following principles to deliver what we call **People Powered AI**:

Principles for People Powered AI	Test
Control. AI should give citizens a clearer and more timely understanding of their health and what should be done, in ways that support greater citizen confidence and control.	Patients should report higher levels of understanding of their condition, control of their health and confidence to manage it.
Simplicity. Well implemented AI should make it quicker and easier for patients to get a resolution to their problem. This requires clarity about the types of problem AI can deal with, and well defined boundaries beyond which human input is required, to avoid AI becoming an additional barrier.	Patients report that it takes less time, fewer steps, and less frustration to get to a resolution of their problem.
Dialogue. The conversation between doctor and patient should remain central. AI should support conversations - ensuring that they are with the right people, that it happens at the right time, and providing the information that supports it. AI should not degrade conversations by over-standardising or taking up unnecessary time.	Patients and professionals should report having higher quality conversations: more time to talk, clearer communication, better mutual understanding and more confidence in the decisions made.
Equity. AI should not be used in ways that exacerbate health inequalities. AI should help all citizens, and most particularly those who face the most challenges and disadvantage in relation to their health and wellbeing.	All previously mentioned metrics, analysed for equity.
Accountability. It must be possible for AI to be understood, questioned and held to account, otherwise AI could fundamentally disempower users - both citizens and health professionals. Without accountability (and the transparency underpinning it), the rest of the People Powered AI principles are hard to achieve - control, simplicity, dialogue and equity all require AI that can be understood and held to account by its users.	Pending European legislation (GDPR) allows for a right to an explanation of a decision from an algorithm. This should be maintained to provide the ability to scrutinise decisions and improve performance.

These are principles that apply to any form of healthcare that aims to be humane and person-centred, but are not presently being applied to the design, development and implementation of AI.

AI development is being driven by private companies, who are not directly incentivised to think from the system point of view. If we sleepwalk into a situation where a small number of tech companies have already monopolised access to the data to build AI, and are selling into a health service which does not fully understand the technology they are buying, the more negative scenarios for AI become much more likely.

There is currently a window of opportunity to shape the future of AI in health. Policymakers should set rules for AI and ownership of public data that ensure the public gets not only value for any data it decides to share, and privacy elsewhere, but also AI products that deliver maximum public benefit. This requires that both the providers and users of AI understand the technology, have the tools to shape the market, can understand the needs of citizens, and are able to work through the complexities of implementation. This can be achieved through the following four recommendations:

Recommendations

1 Public and clinical scrutiny:

Involve citizens and clinical professionals in the upstream design, development and implementation of the technology. This should include the requirement of mechanisms, such as public panels made up of citizens, that ensure technology development and implementation takes account of the demands and perspectives of citizens and healthcare professionals and ensures that People Powered AI principles are applied.

2 Controlled tests in real-world conditions:

Enable real-world experimentation of AI in designated test sites, with non-AI comparators, to understand how AI works in complex systems before wider take-up ‘in the wild’.

3 Proactive market design:

System leaders actively engage in market design to maximise public benefit and ensure a plural market with genuine choice. This should include regulation that is upstream and proactive (‘anticipatory regulation’), clarity over who owns both algorithms and data, and requiring adherence to key design principles, such as People Powered AI principles. Market design should also foster a diversity of new entrants to the market including procurement processes that work for smaller companies and market structures that support a diverse range of R&D activities.

4 Decision-makers equipped to be informed users:

Create a new cadre of public leaders and decision-makers with the technical skills, authority and institutional levers to scrutinise, manage and deploy AI in a responsible way. This should include incorporating artificial intelligence into medical education and health management training to enable the frontline workforce to be informed users of the technology.

Introduction

Machines are getting smarter, and doing so quickly. They can beat humans at complex games, drive cars, and, when presented with the right data, make an accurate clinical diagnosis. Machines can successfully accomplish tasks that, only a few years ago, were the exclusive domain of human beings.

This Artificial Intelligence (AI) seems likely to be a transformative technology in general, but could be particularly significant in healthcare, which is rich in the data on which AI thrives, and the types of problems that it is able to tackle.

AI-driven healthcare has made some impressive technical achievements. A growing body of research shows that AI is capable of performing at a similar level to a doctor across a range of diagnostic tasks. There are already commercially available AI products which have a conversation with a patient and put forward a tentative diagnosis, such as bablyon,[3] and recommend cancer treatments to patients, such as IBM Watson

This success has lead to a degree of hype, including the suggestion that AI will rapidly come to replace doctors. In fact there remain significant constraints on what the current generation of AI can do, a lack of evidence to back up some of the claims made for AI technologies, and significant issues of accountability and trust to be overcome.

But replacing doctors is not the only way AI could be significant in health. An easier route for AI is to provide solutions where few exist, such as advice before we see a doctor, predictive and real-time analytics of health data, or a digital second opinion. AI could also pick up a significant amount of 'lower value' tasks, supporting existing clinical work.

If AI can deliver these new capabilities, then despite being less headline-grabbing than robot doctors, they could really help. For example, they could transform access to care by providing 24/7 advice and triage, and relieving pressure on an overstretched health system.

However, this does require putting AI in a very influential position. For better or worse, AI could change patients' experience of care: how care is accessed, the way services are organised, and the relationship between citizen and clinician. Whether this improves or degrades the experience of care for citizens depends on how the technology is designed and implemented. If the pace of development of AI continues, and given the amount money being invested in it, and the 'burning platform' of an overstretched health service AI could become commonplace in as little five years. As the legal, regulatory and commercial rules that govern AI are being set, we have a window of opportunity to ensure that patients are placed at the centre of the process, and that AI delivers the right outcomes for them.

Consequently, the goal of this report is to:

- Explore how AI might be used in the UK, or a similar health system.
- Explore how AI-enabled healthcare might look and feel, especially from the point of view of the citizen.
- To suggest what can be done to maximise benefits and minimise harm.

We have not sought to make predictions about the future capabilities of AI. Nor do we pretend to cover every variety of AI, or do justice to every issue raised by the technology. Instead we focus on the uses of AI that have the clearest route to adoption, potential to solve real problems, and strong impact on the experience of care.

The remainder of the report is organised in five sections.

1. AI primer
2. How AI might be used
3. How we might experience AI
4. Power and autonomy
5. Conclusion

Box 1 - Super-intelligent machines?

The potential of AI has led to the most enthusiastic commentators suggesting that AI has crossed a rubicon and will now rapidly exceed human intelligence. This has led to stark warnings from Nick Bostrom, Stephen Hawking, Elon Musk and others, of super-intelligences taking over. This is not the first time there have been such ambitious predictions, both utopian and dystopian, about a wholesale replacement of people by machines. The early AI researcher and polymath Herbert Simon wrote, in 1965, that:

> *"Machines will be capable, within 20 years, of doing any work a man can do."*[4]

The developments that Herbert Simon was seeing in the 1950s and 1960s - the ability of machines to solve high school algebra problems described in words, or prove mathematical theorems - did not translate into the ability to, for example, recognise a face. We had to invent and perfect a new technology, decades later, to do that.

Artificial general intelligence, which could perform the full range of tasks that a human mind can, is considered to be unlikely to come from merely improving our present techniques. Most of the experts interviewed for this research believe we will require fundamental, new breakthroughs before we can build machines that can perform as we do, or better.

Consequently this report does not consider artificial general intelligence. We focus instead on the technologies available at the moment, and their likely progression.

1

AI primer

In this section we define what we mean by AI, highlight the varieties of AI considered in this report, and illustrate their achievements and limitations hitherto. Those familiar with the field may wish to skip ahead to the next section.

Defining AI

In this report we use the term artificial intelligence (AI) in its colloquial, informal sense to mean computers which perform cognitive tasks usually associated with human minds, particularly learning and problem-solving.

This lack of precision stems from the fact that there is no definitive understanding of what intelligence means, and what tasks require it. It is this that creates the uncertainty about what the impact of AI might be, and is consequently a question with which we will have to grapple in this report.

In doing so, we limit ourselves to capabilities of AI that do not seem to require a big technical breakthrough; as described in Box 1 above, this excludes AI that can display the same range of abilities as humans. Instead we are looking at extensions of what AI can do at the moment.

Varieties of AI

"AI is not really any single thing - [it is] a set of rich sub-disciplines and methods, vision, perception, speech and dialogue, decisions and planning, robotics"

Eric Horwitz, director of Microsoft Research Labs[5]

The term AI covers a range of distinct approaches and abilities. In order to highlight the potential of AI we will examine three of the most important approaches, which underlie much of the leading research and products at the moment. While there are other AI technologies, these were by far the most frequently mentioned in our interviews.

- **Pure Machine Learning** - ML is when machines learn from examples, rather than being explicitly programmed. Rather than being, in some sense, told what a cat looks like, a machine would learn by being shown many images of cats, and told *"cats are things that look like this"*, or learn to tell cats from dogs by viewing pictures of them both.
- **Chatbots** - A chatbot is a programme that can conduct a conversation, and answer questions requested of it in natural language. Most of us have such a technology embedded into our phones. Beneath the natural language interface, (often built via machine learning) there is a more explicitly organised knowledge base that the chatbot can query to find its answers. This allows chatbots to be broader than pure machine learning in terms of the range of problems they can tackle.

- **IBM Watson** - Like chatbots it uses natural language processing and a rich knowledge base. However, rather than a chat interface, it reads the medical records of an individual, and compares them to understanding drawn from work with doctors, and published research. From this it makes a treatment recommendation for the patient.

Each of these attempt to do a slightly different medical job, and their consequent potential impact on the health system is different too.

Box 2 - Pure Machine Learning

The single most important driver of the improvements in AI over recent years has been the success of Machine Learning (ML) algorithms.

What ML can do

Machine learning has shown real success at making diagnostic judgements from rich data such as images and audio. In specific tasks it is showing ability comparable to that of a trained clinician. Examples include:

- **Opthalmology**. Google DeepMind's work on Diabetic Retinopathy,[6] showing the ability to diagnose this condition from images in a way that is comparable to a trained clinician.
- **Dermatology**. A recent paper in *Nature*[7] showed a machine being as capable as a physician in visually identifying certain types of skin cancer, and there are apps, such as Skin Analytics,[8] which are publicly available, to do this (although they stop short of a definitive diagnosis).
- **Neurology**. Parkinson's' Voice[9] can make a reliable diagnosis of Parkinson's Disease from a recording of a patient's voice, picking up on slight differences in fine motor control.

Word choices for psychological insight, gait analysis for musculoskeletal conditions, accelerometer data for diseases of motion are all active areas of research. There are also companies, at various stages of development, offering ML to support imaging diagnosis in cardiology,[10] neurology,[11] psychology[12] and many other specialisms. At this stage only a few have regulatory approval, but the next few years should see these available in a broad range of areas.

Limitations

A key limitation is that, while ML can produce performance comparable to the best humans, this is for tasks that are relatively constrained. Examining a fixed input, e.g. an image of a particular kind like a brain MRI, to make a binary diagnostic decision, is a heavily constrained problem. A conversation with an older person who has several long-term health conditions, on the other hand, can require synthesising clinical knowledge, psychology, understanding of social norms, local knowledge of treatment pathways, and the ability to have a conversation on a sensitive topic. This leads people to talk about narrow AI - AI that works well for specific tasks on certain kinds of data, but not for broad judgements.

Box 3 - Chatbots for decision advice

Access to a knowledge base allows chatbots to be broader than pure machine learning in terms of the range of problems they can tackle. Chatbots aim to be able to give useful responses in a broad range of situations - up to and including operating as a general source of advice across all conditions. However, they tend to be shallower, in terms of the complexity of the conditions they can deal with and the conviction of their response.

What chatbots can do in health

There is already a wide range of health chatbots. Some are purely informational, answering simple questions in natural language that may be more accessible than a pamphlet or textbook. Arthritis Research UK's 'Arthy'[13] chatbot offers detailed advice aimed at a specific long-term condition. Rather than handling urgent queries, it is designed to answer questions about *"day-to-day life, symptoms and treatment options"*[14] drawing on the charity's extensive body of evidence and questions.

Others are essentially triage, offering a tentative 'diagnosis' and a recommendation for taking action, whether advising self-care or suggesting where to get clinical help. Examples include your.md, Ada and babylon health.

A final class of chatbots support treatment, particularly for mental health. These bots draw on cognitive behavioural therapy (CBT), a highly structured, repetitive treatment intended to change negative patterns of thought. Woebot, a tool which helps identify and shift negative patterns of thought, *"significantly reduced"* symptoms of depression in trials.[15]

Limitations

Unlike some pure machine learning approaches, chatbots are not yet well evidenced, except in mental health, where there is a small but growing body of research. There is a set of unanswered questions about their limitations.

Can they give safe advice? Where they are purely offering information, and operating as a more interactive and responsive version of sites such as NHS choices, AI simply needs to be better than the alternative. But where they operate as a form of advice and triage, there are significant safety concerns, and a need for rigorous evidence. In some cases there are concerns about them being rolled out with insufficient evidence.[16]

Can they avoid producing large numbers of worried well. Chatbots that overreact to symptoms and send large numbers of people into the health service would be destabilising to the system and unhelpful for users.

Box 4 - IBM Watson

IBM Watson is something of a special case - there is no other product that offers similar functionality, or so squarely competes with doctors in its capabilities. As it offers treatment recommendations rather than just a diagnosis, Watson could be used to replace a consultant - and indeed Watson is being used at UB Songdo Hospital in Mongolia without any oncologist supervision. However, more typically, its use is advisory, or consulted when there is a difference of opinion on the team, functioning as a supportive technology

Oncology is the first specialism that IBM has developed. IBM quotes a study from Bangalore in which the system's recommendations matched the local multidisciplinary tumour team's ones 96 per cent of the time for lung cancer, 93 per cent for rectal cancer and 81 per cent for colon cancer. However, this research has IBM employees as co-authors. There is a lack of independent and peer-reviewed research on how well Watson performs.

Limitations

Watson has a number of limitations. It is trained specifically on wealthy US patients by clinicians using US guidelines. In South Korea clinicians could not follow Watson's recommendations because it requires drugs not covered by local health insurance. It is also hard to keep up to date with rapidly changing guidelines, as it requires a high degree of manual programming.

The lack of independent research on Watson Oncology makes it hard to be confident about the quality of the product, or to fully understand its limitations.

There has also been a certain amount of unhelpful marketing of Watson, which was taken to suggest that it could automatically update its recommendations as new research was published. This is not the case with the Watson Oncology product.

Data challenges

Practitioners often focus on availability of data as the main limitation to the progress of ML. Machine learning requires enormous quantities of data to tease out patterns and cluster data according to common properties. Generally speaking, the more data, the better the model that will be developed, the more useful patterns might be uncovered and the better predictions can be made.

The availability of data is, at present, far from where ML developers would like it to be:

- Data protection legislation and norms makes accessing many large datasets a challenge.
- Data is often not of sufficient quality. Frequently the data is incomplete, and with a significant number of errors, fragmented between different repositories, and hard to link up accurately, or coded using incompatible frameworks.
- Data that ML developers want may simply not exist, either because it has never been digitised, or because it has never been recorded at all. A lot of potentially valuable data collected by devices, including monitoring equipment, is immediately discarded after use or display. And a lot of information that doctors use in practice never makes it to the record.
- The data may not have been gathered in a way that makes it possible to distinguish causation from correlation - see below for an example.

> *"A lot of information that doctors use for decisions isn't captured, so it isn't there for AI to learn from. Doctors could pull in information about social situation, context. If you're treating a frail, elderly person with no social support you make a different decision at 2am and 2pm."*
>
> **Alastair Pickering, Clinical Lead (Urgent and Emergency Care) NHS Digital**

All of this means preparing data for machine learning applications requires significant amounts of technology, investment, and trained human labour. Data has often been called the new oil; it should not be surprising that alongside collecting, this material also needs refining, and that this refining process requires skill and investment.

The speed of progress towards the futures sketched in this report will, to a large extent, be determined by how we solve this set of problems.

Data complications

A team at the University of Pittsburgh Medical Center wanted to better predict which pneumonia patients would develop severe complications, and so be able to manage demand and improve patient safety; sending low-risk patients to outpatient services, whilst better identifying those at high risk and admitting them to hospital. They tried both neural nets and traditional methods such as logistic regression- 'rule-based systems'. They found that neural nets outperformed other methods by a wide margin. But there was a downside.

It was hospital policy to send people with both asthma and pneumonia not only to an inpatient bed but directly to intensive care, as the combination of the two can produce severe complications. This policy worked so well that people with asthma almost never developed these complications.[17]

In learning from this dataset in isolation, the AI came to the conclusion that having asthma reduced the risks associated with pneumonia, the exact opposite of the truth. It recommended community management for these patients, a dangerous course of action.[18] The AI was picking up a pattern that did exist in the training data, but which did not reflect reality.

The team used both an AI with explicit and visible rules, and a neural net where the logic is much harder for a human to understand. This failure was visible in the rules-based system, and so easy to catch. But the study authors conclude *"If the rule-based system had learned that asthma lowers risk, certainly the neural nets had learned it, too."* If the team had relied only on a neural net and trusted its conclusions it could have made deadly errors.

2

How AI might be used

In this section we consider how the present generation of AI might be adopted within the healthcare system, and hence be part of our everyday experience of healthcare. We argue that AI will see its fastest adoption in areas where it is not competing head-to-head with humans, but even so will be in a hugely influential position.

The present generation of AI is narrow compared to people. It works for tasks that are to some degree constrained. An ML algorithm is generally trained on data of a particular type (e.g. an MRI), and learns to reach certain types of conclusions (e.g. cancer present or not). And while chabots are capable of covering a broader range of situations, the conversations they can have are still relatively simple.

In contrast, much of medicine requires complex and holistic judgements. It requires a clinician to use what is effectively a general intelligence, integrating a whole range of capabilities to make a judgement - linguistic, interpersonal, physiological, clinical, and psychological understanding as well as knowledge of the local landscape. As mentioned in Box 1, it would probably take a fundamental breakthrough in AI for machines to be able to do this. Indeed, in a world where AI could do this, it could probably write this report. Consequently most people think that large sections of medicine will remain hard for the current generation of AI.

It is also fair to say that there are significant areas of medicine that involve narrow application of decision-making rules and expert pattern matching. Diagnostic specialisms such as radiology and pathology are seen as containing a significant number of tasks in this area - although there is more to these specialisms than that.

It is an open debate how much of medicine can be reduced to narrow judgements based on a consistent set of inputs, and how much requires a holistic and general intelligence.

But even when AI is successful at these more constrained tasks, there are obstacles to machines replacing people completely. To actually replace doctors, AI has to show superiority in nearly all circumstances - or equivalently, that the doctor adds almost no value in combination with the AI. In practice, machines and people have different strengths. For example people are better at learning from a smaller number of examples, and therefore are likely to do better in rare circumstances - unusual presentations, rare diseases, etc. AI needs large amounts of data to learn to perform a task, which may be hard to assemble for rare conditions and circumstances. People also have common sense, and can spot when the machine is making an obvious mistake due, for example, to a foreign object that has found its way into the medical image.

Additionally, the public prize their personal relationships with their doctors, and have a high degree of trust in the medical profession. And having a human in charge leads to clear accountability for decisions. The public are likely to set a very high bar when it comes to anything that might undermine these, and with good reason.

Generally, competing head-to-head with humans at tasks that people are already good at, is a difficult route to adoption. Easier routes are likely to be: helping people to perform their existing tasks more easily, and providing new solutions where there are presently no good alternatives.

AI can take lower value tasks that people do not want to perform. AI could, for example, help a pathologist by counting the number of abnormal cells in a sample - a task that is both dull and difficult for people to be accurate at. This is likely to be less disruptive to relationships, and so is less relevant to the focus of this report.

The other, and perhaps the easiest route for AI to come into our lives, will be offering solutions where is there is significant need but presently few good alternatives on offer. Innovations often take root in this way, as argued by the great innovation theorist Clayton Christensen.[19] We term this **Complementary AI**. Areas that caught the interest of our interviewees included:

Advice and triage before seeing a doctor

When should someone first seek help from the health service?

Proactive care

When is the right time to intervene in the face of worsening symptoms?

Automated second opinion

How does the diagnosis and treatment I am getting compare to alternative options?

There are AI driven products in all these categories, although not always with sufficient evidence of safety or efficacy as yet. These products could be very attractive to both an overstretched health service, and to worried citizens. This would put AI in a hugely influential position for our health, our experience of care, and relationships within the system. The following paragraphs explore what this might look like.

Potential impacts of complementary AI

1. Advice and triage before seeing a doctor

One area where existing solutions are weak is before the patient chooses to first contact a doctor - what some call the 'pre-primary care'[20] space. Online advice is often unreliable, contradictory, and not specific enough to be helpful in individual circumstances.

With AI, members of the public could have access to high quality screening and personalised advice. Chatbots are available that answer medical queries for a range of situations, such as babylon health, Ada and your.md. And smartphone apps can make AI imaging analysis easily available e.g. for skin cancer[21] - although app advice often stops short of a definitive diagnosis. As well as helping with self-care and diagnosis, these products will inevitably give advice on when to seek medical care - effectively triaging people into the health system.

This could be very significant. The fact that it is hard for patients to know when to seek help means A&E and GP practices are full of people who did not need to be there (but who largely could not have known that at the time). Research suggests that 20 per cent of GP[22] appointments and 19 per cent of A&E attendances[23] are for more minor medical problems that could be treated at home. Equally, there are people who do not seek medical attention when they should, often causing further problems and expense at a later date. Progress on these points would be hugely significant to the economics of the health system, as well as quality of care.

There are also significant risks. The AI could give unsafe advice. It would also be easy to generate a flood of worried well, should the AI be too risk averse in its recommendations and throw off a large number of false positives. With the potential legal and reputational risks of unsafe advice, there will be significant pressure on AI developers to be cautious. Widespread consumer adoption of overly risk averse triage AI would risk overloading a health service already at breaking point.

Where there is explicit or implicit pressure to use the AI, it could also easily become a barrier to be negotiated, rather than an aid. Opaque reasoning, and a failure to cope well with complex or atypical cases, would lead to frustration and perhaps gaming of the system.

AI advice is already being rolled out here in London as part of the babylon health's GP at Hand service.[24] This combines remote access to a GP with a triage chatbot, and is available to anyone who lives or works in central London. The service has received some criticism of is lack of an independent evidence base, and a review is underway.[25] But in the near future data will emerge that establishes whether the advice given is safe, and whether it meets demand appropriately. If this service can safely reduce unnecessary demand, then it would seem likely that it will be rolled out at scale in coming years.

2. Proactive care

Similarly to the dilemma of care, it is hard for patients to know themselves when to seek help for a potentially deteriorating condition, and expensive for the health service to frequently assess them. Sensors and devices that can collect data from the patient, and trigger more timely care - known as telehealth - have been around for many years. However, this technology has not quite broken through to the mainstream, despite considerable effort and investment from government. AI brings new analytical capabilities that could hugely expand the scope and quality of this real-time monitoring.

For example, an app from Cordio can hear signs of deterioration in patients with congestive heart failure, via an audio recording of their voice. The app detects signs of a build up of fluids around the patient's lung, before other physical symptoms are manifested.

Technologies of this kind should allow earlier intervention, and so improve outcomes and reduce costs. They would reduce unnecessary appointments, while responding more quickly and effectively to problems. It would require a significant rearrangement of staff and modes of communication between them, but the preventative savings and improved outcomes from timely intervention could be significant. Again, if this sort of technology can be established as reliable and safe, then it could be a significant part of our experience of care in coming years.

There are further risks beyond safety and reliability. Overly risk averse technologies could create overload on the system, as well as patients who ignore the warnings. Ubiquitous monitoring could easily feel excessive and even nagging. Finally the system would be data hungry, and filling in gaps in the data could easily reduce time for conversation.

3. Automated second opinion

Healthcare is one of the interactions in which we are least able to question and assess the advice we receive. AI could change this.

AI could be used to as a comparator or digital second opinion. Technology like IBM Watson (if proven to be reliable) allows a whole treatment approach to be compared to an automated alternative. The recommendations of local clinicians could, in theory, be compared to those of AI trained and programmed at the most highly respected hospitals - perhaps the Royal Marsden approach to cancer, or a Moorfields approach to eye problems.

Managers and regulators could also have this option - perhaps using it to screen for sources of variation in treatment. If AI reaches high levels of accuracy for certain situations, a doctor who gives very different diagnoses regularly might raise questions.

To be clear, the evidence certainly does not support making AI a standard to which doctors should be held, even in limited circumstances. But that will not prevent individuals and organisations with access to AI making comparisons. Patients find it hard to know if they are getting appropriate treatment; and given the importance of the issue, an easily accessible second opinion is appealing. Similarly managers and clinical leaders both seek to understand and control variation in treatment and outcome for similar patients.

Where the AI is accurate and applied within its competence, and the results are understood in context, this could be helpful. However, again, AI could easily be highly problematic in this context if misapplied. Inaccurate AI advice could undermine the trust between doctor and patient. Central monitoring could put pressure on clinicians

Overall, AI in these three categories:

- Could be very attractive to patients and managers, potentially solving problems that are serious and significant.
- Carries with it a great deal of responsibility, not only in terms of safety, accuracy and efficacy, but also in its impact on power and control for both patients and professionals.

Chapters 3 and 4 explore the impact of this sort of AI from the perspective of patients and professionals.

	More power	Less power
Advice and triage before seeing a doctor	Patients finally have a sound basis for seeking help or not, and are in a much stronger position to make that choice for themselves. They are better informed for the consultation that follows, with a tentative diagnosis that they have the opportunity to research.	Implicitly or explicitly, patients are compelled to complete an AI triage to access care. Its reasoning is opaque, and it fails to cope well with complex or atypical cases, leading to frustration and gaming of the system. This gaming undermines the doctor-patient relationship.
Proactive care	Patients with long-term conditions can understand what is happening to them in real time, and get help when they need it. The system is more responsive to patients' needs, more efficient and timely.	Monitoring feels excessive and nagging. Compliance is poor, and plugging gaps in the data squeezes time for conversation. Variation between patients is underestimated, leading to inappropriate responses from the AI, wasted resources and unhappy patients.
Automated second opinion	Patients can access an alternative treatment plan for their circumstances - this prompts a much more equal conversation between doctor and patient about options, and eliminates some poor care.	Central monitoring of decision-making via an AI that is beyond its competence could create inappropriate standardisation, and less account taken of the individual patient, and aspects of the locality.

Box 5 - Superior AI - how could AI displace doctors?

Our point above is that the fastest and most likely way that AI will come into all our lives is by solving problems that do not have a good solution, or helping doctors to do what they already do, rather than competing head-to-head on things they are already good at. However, this is not to say that AI cannot replace people - simply that it has to be unarguably superior.

Often, this would mean AI that generated what is known as stratified or personalised treatment recommendations. Underneath an identical diagnosis, there are huge amounts of variation between individuals. Two people who have had a stroke can have quite different patterns of damage in their brains, and quite different prognoses. While we classify cancers by their size and location, there is a great deal of underlying genetic variation in the cancer, and this is believed to be partly responsible for different outcomes. And on top of this underlying difference in the causes of the disease, there is of course variation in age, physical condition, and other diseases, all of which affect response to treatment and outcomes.

If we could understand this variation, we could design tailored treatments, based on individual factors for individuals, that would bemore effective than at present. This is referred to as stratified medicine, or in its most ambitious version, precision medicine. ML, with its ability to find complex relationships between large numbers of factors, could be the technology that enables it. A live example is research underway at UCL to decide which stroke patients would benefit from speech and language therapy, by looking at MRI scans. This would help make sure a scarce resource went to the right people.

Critically, in the individual case, humans could not check the logic of the decision made by the AI. The ability of ML to find and take account of a very large number of factors in its calculations is part of its power, but this makes the decision potentially opaque, and often not verifiable by a human. This puts humans outside the decision-making loop. While we would still interrogate and accredit the technology centrally, local clinicians would not really be able to overrule the AI in most circumstances, as they have no way of understanding if it was right or wrong, and no alternative route to the same judgement. At this point the job of a clinician becomes very different - perhaps operating as more of a navigator and advocate for the patient than a decision-maker.

It is important to note here that there is no guarantee that this sort of insight is possible, with our present level of scientific understanding. Twenty years ago it was widely expected that genomic data would allow us to make individually personalised treatment recommendations, especially for drugs. This has not happened at any significant scale. Often we do not understand the underlying mechanisms of a disease well enough to know which data would allow us to personalise treatment. And there is no guarantee that we will find it. However, this is a very active area of research.

3

How we might experience AI

In this section we will try to be more concrete and specific about how AI might change the experience of healthcare for patients and professionals. We will do this via a series of illustrative narratives.

Scenario 1 Power reduced by AI

Omar has taken up his employer's offer of a free health tracking wearable and app as part of a wellness programme. Omar hopes the wearable might give him answers for why he feels more fatigued these days, and often has aching joints after a long shift. At 39 years old he is unsure if this is a normal part of ageing; he doesn't want to complain but the pain is often unbearable.

He has read about precisely tailored exercise regimes that learn from detailed movement data, but can't afford the subscription to these services.

When the company wearable turns up Omar is frustrated that he cannot access the raw data and look at patterns, but instead is only shown how close he is to reaching company-set exercise, sleep and diet goals. The wellness programme keeps recommending him more exercise, and deducting points from his score when he fails to complete it, but Omar's pain and fatigue is getting worse, and he can barely keep up with work and childcare.

On a lunch break Omar uses an NHS triage app to try to get an appointment for tailored advice, but he struggles to articulate the connection between his symptoms; his joint pain has become so normal that he attributes it to work, and he forgets previous symptoms. The app advises rest, an exercise routine, and iron supplements for mouth ulcers and possible anaemia.

Increasingly desperate, he follows a friend's advice on gaming the system, using specific keywords in the triage app, including 'stabbing pains'. After a few attempts Omar gets an urgent appointment at an oversubscribed clinic. The doctor is not happy to find that Omar has lied, and marks his record for gaming the system. The clinic is under pressure to prevent gaming and lower costs.

Suspicious and busy, the clinician is guided by a strict script, imposed to allow AI analysis of decision-making and reduce variation. Dwelling on fatigue and pain, and noting Omar's warehouse job, he follows the system recommendation of gradually increasing exercise to build strength, and iron supplements for possible anaemia. He does not have access to the company wearables' data stream so cannot see the longer-term patterns in symptoms. He notes that Omar looks very well, which increases his frustration and suspicion. But there is no place to enter this detail in the system - which might otherwise have flagged this 'glow' as a classic sign of autoimmune problems.

At home, low on shifts, Omar researches his symptoms himself, and his posts about pain and fatigue trigger adverts for Lupus groups. Omar is lucky, and finds a genuine peer support group sharing stories that reflect his own experience. Interested, he inputs the query into the NHS triage app, but being unfamiliar with the terms used he fails to input symptoms clearly. His previous use has him flagged as problematic, and the app suggests he adhere to the previous exercise regime and supplements.

Finally, after discussions on a Lupus Forum, Omar crowdfunds to pay for a private appointment, to pay for access to the data on his company-owned wearable. Using this data, the private clinician diagnoses him with Lupus. The diagnostic process has taken five years; as long as it took on average in 2017.

In this dystopian scenario we see a number of the potential risks of AI in a real implementation.

Failure to support control

We can easily enter into an era of 'big brother' medicine, where patients' symptoms, diets, exercise and medication adherence are excessively monitored and controlled.

Little effort is made to understand how the data can be made useful to patients, rather than to the system. Insufficient attention is paid to ensuring the patient is sufficiently engaged to make the choices that are right for them, and supporting their autonomy and control over their health. Rights to transparency, explanation and to challenge decisions are weak.

Critically AI is used beyond its competence in this example, with no mechanisms for patients or clinicians to override its recommendations. In this case the AI is not sophisticated enough to elicit some of the more subtle but relevant symptoms, and is not implemented in a way that is compensated for. Gaming is an inevitable consequence, and a degree of suspicion is introduced to the doctor-patient conversation, weakening a critical relationship.

Degraded dialogue

The presence of a rigid data-driven script disempowers both doctor and patient, draining their interaction of autonomy and clarity. Implementation pays insufficient attention the role of the conversation between doctor and patient in eliciting clinically relevant detail.

Scenario 2 Power increased

Selina is 64, living just outside Norwich. She had breast cancer in her thirties, so is now on a frequent screening programme. She pops into a supermarket pharmacy and has a quick puff on a device which collects biomarkers in breath. These use deep neural networks to identify biomarkers for early detection of many cancers, as well as lung, liver and kidney conditions.

The system identifies markers of very early-stage colon cancer immediately but because of the uncertainty level, does not offer a specific diagnosis. The screening technician simply lets Selina know she needs additional tests. At the appointment, a scan identifies a possible lump and a biopsy is taken for genetic sequencing.

At the news of a cancer diagnosis, Selina is devastated; her first experience of cancer was extremely traumatic; she still feels pain in the surgical scars, and recalls the gruelling year of chemotherapy. A nurse explains that early identification and new technologies mean treatment should be short and targeted, with an excellent prognosis, but Selina is too upset to absorb the details. She takes a recommendation for support services and goes home.

Over the next week in the comfort of her own home Selina tries a number of different apps that are all built on the NHS's accredited cancer chatbot platform. She finds that she prefers one from Macmillan. Alongside access to the latest medical advice in natural language, many answers are supplemented by notes of advice or anecdotes of other patients' stories. It also allows her to dial up or down the level of detail in answers; at first she reads brief summaries, but as her anxiety fades she reads in more and more detail, occasionally reading the latest medical studies and looking at open data resources.

Selina's chemotherapy is targeted specifically to the vulnerabilities of her cancer; an oncologist works with an artificial intelligence which identified potential drug cocktails. Because they are so precisely targeted, the predictions do not have a lot of totally similar patients to learn from, so the final decision is a collaboration between the oncologist's experienced opinion, Selina's tolerance of risk and side effects, and the AI.

One night, Selina feels too nauseous to sleep, but being exhausted, she feels anxious about bothering a nurse, or waking up her daughter or a friend. She opens the app and explains her symptoms; she is glad that with a bot she can take her own time to gather her thoughts, frame her questions, and repeat answers.

The bot, which has access to her medication records, confirms that nausea is a recognised side-effect but that it is safe to rest at home for now and visit the doctor tomorrow morning to receive an updated anti-nausea prescription. Alongside a simple button press to confirm the appointment, the app includes a short anecdote written by another patient explaining how horrible the sickness felt, and how glad she was that a new drug had helped. There is a link to a patient discussion forum, staffed by volunteer moderators all night.

Still unable to sleep, Selina investigates statistics on side effects, eventually ending up on a patient-led quality checking forum. They have used open source regression AI to look at patterns across open care quality data. Selina enters in demographic details and is surprised to find that her cancer specialist clinic is much worse than average at managing side effects in older women. The system suggests that new procedures that use mathematical tools to check for bias in decision-making have not been implemented, and that older women's pain is less well monitored and medicated. At her next appointment she raises this issue, and the team offer to present her data at a procedure review meeting. The supplier of the oncology software is asked to review its algorithm for bias and after investigation is found to have used a historical dataset with problematic biases. The system is updated. Although the process takes only a few months, by the time she hears the outcome of her campaign, Selina has already completed her targeted treatment and been declared cancer free.

This future healthcare system has made implementation choices that support a powerful patient. In particular AI is used judiciously and in combination with other sources of advice and insight, each of which has their place. This results in a positive outcome along a range of dimensions.

Control

AI tools are designed to maximise control. They allow Selina to get a powerful screening test done in a convenient way, access advice that she can tailor to her needs, and check up on the performance of her clinicians. Knowledge and expertise are more available to her than at present, and are delivered when and how she wants them.

Simplicity

The AI is applied with a degree of realism, not being asked to do more than it can. Issues of the reliability and completeness of AI advice are well understood, and designed for. Face-to-face and peer-to-peer are used when they are the best approach, resulting in a simpler, cleaner experience for Selina.

Dialogue

A clinical conversation with broad scope remains the centre of the care process, and the AI serves to enhance Selina's power in those conversations.

Transparency and equity

The system is transparent about performance, and open to patient influence to identify and eliminate bias.

Scenario 3 Shifting the GP role

Harpreet begins her day by reviewing the work done by the practice AI over the last 24 hours. Over 30 minutes she approves a few dozen suggestions that the AI has made to patients and books some brief phone calls where she wants some more information. Then she turns her attention to the more complex chronic patients who are her main focus.

She sees fewer patients these days, and has more time for each one. From the peak of 40-60 patients a day in 2010, the number has been reduced to a maximum of 25 with this time managed more intelligently. Sometimes this means an AI arranging the most efficient routes and schedules for home visits or prioritising calls and monitoring tasks, sometimes more prosaically, such as group appointments.

The first patient is Carly. The appointment was recommended by Carly's care AI, which had identified worrying trends in adherence and symptoms. A 20-minute appointment has been scheduled, now the standard length for complex patients.

Carly's wearable and self-reporting symptom tracker indicates an increase in frequency of painful episodes in her Crohn's and arthritis. Harpreet can see where Carly has manually tagged a few episodes and entered supplementary comments; a period with no data on intestinal function is where she had been ill and could not keep the swallowable sensor down. She is losing weight; though still within a healthy BMI range, the AI flags the speed of this trend as warranting attention. The system reminds Harpreet to look for indications of stress, depression and isolation.

Carly turns up a few minutes late. She is frustrated; her Crohn's is not responding to medication. Like 25 per cent of Crohn's sufferers Carly also has arthritis; this too is worsening, and the pain makes it difficult to prepare the small, regular, specialised meals that an AI has recommended. A few questions about changes in life circumstances reveal that Carly's relationship has recently broken down; the stress is worsening her conditions, and confusing her schedule, which leads to skipped meals and pills. Harpreet can see immediately that automated reminders from an AI are not going to be enough to keep Carly engaged in managing her health

It took quite some time to convince her to start using the swallowed sensors to monitor her gut health in more detail, and while the data has been useful to Harpreet in identifying the exact extent of the damage, Carly is less convinced: *"but I told you where it hurt!"*

The wearable also reports that Carly is not taking her medication at the same time each day. The pattern is not straightforward, but an AI mining the data means Harpreet is able to show how Carly's patterns of severe pain map to times when in the previous days she had skipped a dose or mistimed another. In each case, Carly explains this was due to pain, or delays in transportation and chores which confused her schedule.

She explains to Carly that the system has identified a cohort of 'people like you' with both Crohn's and arthritis, living in similar semi-rural areas with poor transport links. An experiment is being run regarding a new meal and medication support service. As an option over time the service will gather data and run diet experiments; testing how different ingredients trigger symptoms, while also managing her calorie intake and medication schedule. Carly agrees initially because of the convenience, and consents to the experimental element.

By the end of the 20-minute appointment Carly has relaxed and opened up about her current challenges. She sees the value in more information, and has been able to choose services which will reduce stress at a difficult time. As well as recommending several reminder services to help with medication adherence Harpreet convinces Carly to try attending peer support groups for people with arthritis. For some challenges, empathy and human connections are irreplaceable.

In this scenario, a number of factors support citizen and clinician working well together:

Simplicity

In this example the AI spots the problem and books an appointment for good reasons. The system is designed to hand over from AI to clinician appropriately. The information that Carly gets is helpful and timely.

Control

Monitoring is a big part of the treatment, but it is done as part of an ongoing dialogue with the patient, and is made useful to her.

Dialogue

There is a lot of complexity to Carly's case, and a need for an extended conversation to understand it. In this example, the AI is designed to support the conversation between Carly and Harpreet, rather than replace it - they discuss the data together.

Scenario 4 De-skilling the workforce

Daniel retired from clinical practice as a pulmonologist nearly a decade ago, but still spends a couple of afternoons a week working as a clinical auditor from home. He is shown a series of images and readings and given the option to either verify the AI's decision or send it back to be queried. He found the work interesting at first, as he was sent complex cases the AI could not understand, but in the last year he has felt increasingly like a rubber stamp; simply approving decisions he cannot understand. The records he is shown contain less and less information interpretable by humans, he is given very little time for each decision. The cases contain patients' notes, and records of which apps and services they have connected to their health record, some of which recommend treatment plans.

He is presented with enormous amounts of data and a final decision from the AI, but the algorithm itself is protected by intellectual property laws, and he cannot follow its logic. Rather than a tool to assist in identifying and interpreting patterns, supporting human judgement, the AI logic is impenetrable. His own job signing off the decisions is part of a deliberate attempt to sidestep laws requiring algorithms be explainable.

There have been advances which Daniel thinks are fantastic; apps which listen to the breathing of patients and can identify developing issues in lung function, and early detection of congestive heart failure simply through sound. But he worries about the lack of regulation. Daniel is deeply troubled by the flood of symptom monitoring apps on the market for COPD for several reasons. He is increasingly suspicious that many apps are recommending inappropriate treatment options. As he only has access to individual records and decisions it is difficult to clearly define the pattern, but many patients are arguing for medicines, supplements, services or equipment which he does not really think could have the benefits they are promising. The other key reason is that COPD exacerbations are worsened by stress or panic; a symptom monitor which identifies worrying trends and displays them in an insensitive way could have a serious negative impact on someone's health. He worries some apps seem designed to cause panic and urge patients to demand unnecessary expensive drugs.

Daniel is concerned as he gets older about the possibility of accessing high quality care- there has been a lot less investment in educating medical professionals in recent years which has resulted in centralised decisions, and de-skilling. His nephew Segun dropped out of medical school early due to money worries, and took a shorter 'physician's assistant' course. Segun's children are both care workers and are given very little scope to influence decisions which impact the people in their care, one rushing between short in-home appointments, and another at a residential home where robots outnumber human assistants three to one. They complain to Daniel they never have time to fix underlying problems, and it is difficult to support better adherence when decisions cannot be explained.

Lack of transparency

In this future we begin with a lack of transparency in AI decision-making, with Daniel having no clear idea about the reasoning behind the AI decision, and an implementation of AI in a way that maximises the volume of care without prioritising the value for patients, with little thought to their control. Lack of transparency makes it hard to see if AI advice is safe and efficacious at the individual level.

Degraded dialogue

Clinicians find themselves far less able to have face-to-face interactions, despite being the most appropriate solution for the patient in many circumstances.

In consequence, patients experience far less simplicity. Healthcare is even more impenetrable for the patient than it is at the moment.

4

Lessons for People Powered AI

In this section we pick out some of the consequences of AI mentioned in the scenarios above. We argue that the technology allows for a broad range of options, from very empowered to very disempowered, and that choices made in implementation decide where in that range we arrive.

The impact of AI in the scenarios above varies along a number of dimensions - typical with either positive or negative consequences for the power and autonomy of individuals.

Simplicity and complexity

We have described an AI that often sits between citizens and professions - triaging, alerting, prioritising etc. This means it can make accessing care feel simpler, or more difficult. Talking to an AI chatbot that does not understand an unusual presentation, or non-standard phrasing, or take into account multiple conditions, could be immensely frustrating. It could result in people circumventing or gaming the AI; in turn, this could undermine clinical relationships

A core issue is the degree to which the capabilities of the technology match with the task that it is being asked to perform. In particular, are we asking narrow AI to tackle the complex problems that require a much more general intelligence? This is bound to result in poor advice and frustration. AI needs to be implemented in a way that recognises the scope of its own competence, and that hands over to a human when this scope is exceeded.

Control

AI could give patients more control over their health. Real-time monitoring could allow patients to have a far better understanding of their condition, how it was progressing, and if they were on track. They would be better able to ask for care when they needed it. Or it could reduce the control that patients have. Obscure and hard to question AI decision-making would make it hard to exercise control.

A key here is whether the monitoring is designed for patients as much as clinicians or managers. If its outputs are made as accessible as possible for patients, it will maximise opportunities for understanding and self-care.

Dialogue

High quality conversations should remain at the heart of healthcare. It would be easy to see the time for an individual conversation squeezed out by the time taken to get the AI the data it needs, or dominated by the AI's framing of a problem. The potential of AI to monitor clinician behaviour, if not done very carefully, could result in increasingly scripted interactions that can miss important facts about an individual, prevent flexible, creative solutions, and feel mechanical. Equally AI could create more time for conversation, enrich it with helpful analytics, and ensure that doctor and patient were as well prepared as possible.

Accountability

Certain forms of AI are so complicated that it can be impossible to understand how decisions are made and to check the logic which led to the decision. Some machine learning tools are particularly problematic. They may use hundreds of inputs, assign thousands of different weightings to those inputs and combinations of inputs, and arrive at a model that is impossible for a human to scrutinise.

A system that is obscure in its decision-making makes it hard for the patient or the professional to engage with or question it. The system also becomes unresponsive and slow to learn, as the reason for mistakes is not identified. A lack of accountability and transparency exacerbates the other issues mentioned in this section - it makes any deficiencies hard to identify and challenge.

A key here is that AI remains explicable and interrogable, at least to a similar degree to an expert, and that there is clear accountability for rectifying mistakes and learning from them.

Bias and inequality

Algorithms that give advice are vulnerable to bias just as people are. A well publicised example occured in an algorithm used for assessing the risk of reoffending within the US criminal justice system.[26] This algorithm influenced bail, sentencing and parole decisions, so was of great importance. When the algorithm made mistakes, it did so by overstating the reoffending risk for black people, and understating it for white people. This was despite the fact that the algorithm was not told who was black and who was white.

If we move into a world where AI is advising on who would benefit most from treatment, there is real a risk that inappropriate factors are picked up and bias is replicated and reinforced. For example, those from disadvantaged backgrounds may benefit less from a given treatment, due to lack of resources, social networks or extra stressors. A fair approach might be to give them the extra support they need to get the full benefit from treatment. But since AI can be something of a black box, it may not be clear that the drivers of poor response to treatment are factors that should not be considered, or require a different response.

In all of this, it is important to consider the dynamics of the marketplace. AI development is being driven by private companies, who are not directly incentivised to think from the system point of view and bear the points above in mind. Some companies will be incentivised, by reputational risk, to give conservative advice, even if this might increase demand. They will not necessarily consider the impact of their innovation on dialogue, or understand whether, in the context of a complex system, it makes accessing care simpler or harder. This means that the public sector has to be clear about what it wants from AI, understanding the strengths and limitations of the technology, and actively managing the wider impacts. The window of opportunity to do this is limited. If we sleepwalk into a situation where a small number of tech companies have already monopolised access to the data to build AI, and are selling into a health service which does not fully understand the technology they are buying, the capacity for public influence of AI will be much reduced.

A number of principles and metrics could guide us down the right road here.

Principles for People Powered AI	Test
Control. AI should give citizens a clearer and more timely understanding of their health and what should be done, in ways that support greater citizen confidence and control.	Patients should report higher levels of understanding of their condition, control of their health and confidence to manage it.
Simplicity. Well implemented AI should make it quicker and easier for patients to get a resolution to their problem. This requires clarity about the types of problem AI can deal with, and well defined boundaries beyond which human input is required, to avoid AI becoming an additional barrier.	Patients report that it takes less time, fewer steps, and less frustration to get to a resolution of their problem.
Dialogue. The conversation between doctor and patient should remain central. AI should support conversations - ensuring that they are with the right people, that it happens at the right time, and providing the information that supports it. AI should not degrade conversations by over-standardising or taking up unnecessary time.	Patients and professionals should report having higher quality conversations: more time to talk, clearer communication, better mutual understanding and more confidence in the decisions made.
Equity. AI should not be used in ways that exacerbate health inequalities. AI should help all citizens, and most particularly those who face the most challenges and disadvantage in relation to their health and wellbeing.	All previously mentioned metrics, analysed for equity.
Accountability. It must be possible for AI to be understood, questioned and held to account, otherwise AI could fundamentally disempower users - both citizens and health professionals. Without accountability (and the transparency underpinning it), the rest of the People Powered AI principles are hard to achieve - control , simplicity, dialogue and equity all require AI that can be understood and held to account by its users.	Pending European legislation (GDPR) allows for a right to an explanation of a decision from an algorithm. This should be maintained to provide the ability to scrutinise decisions and improve performance.

Box 6 - Value, ownership and access

In a world where a more convenient way of ordering a taxi is worth $50 billion, the rewards for the real clinical breakthroughs that health AI could bring could be huge.

The key asset will be access and control of data. As mentioned elsewhere, most developers feel they have the tools to make powerful AI, but not the underlying data. Large tech firms are engaged in a scramble for data, signing contracts all over the world to gain access. Elsewhere, Nesta has predicted that one of the big global tech firms will buy a major healthcare provider.

This means that health data, which the general public never sold to anyone, could generate vast fortunes for others, based on charging consumers and taxpayers. This generates questions about who benefits from the use of health data and, in particular, how financial value is shared and distributed between parties. The NHS is in possession of some of the largest health data sets in the world, but these are lower quality than generally believed. However, the NHS could set out to build large and high quality datasets, and use this to support a public stake in the rewards of AI.

However, doing this requires either rewriting the rules about how people exercise control over their own data - a sort of data nationalisation - or working creatively to find new ways of engaging with patients as active participants in what happens to their data via, for example, new forms of collective governance, or group-based ethics decisions.

Previous attempts to share NHS data, such as the care.data project, have resulted in public resistance. This means that data governance and ethics are viewed as a third rail, and there is little sign of any major organisation wanting to take this on.

But even if the public sector does not act, the private sector will. It will build datasets through partnership agreements, direct purchases of healthcare providers and insurers, use of wearables and sensors, and any other means at its disposal. And will consequently accrue the benefits.

Box 7 - Privacy

We have mainly been concerned with the potential impact of AI in use. However, the creation of AI also has implications, particularly for privacy.

Machine learning requires a lot of data to train, and in health this data is often personal and sensitive. While obviously identifying data can be removed, this is not a perfect protection. Eighty-seven per cent of Americans can be identified using only their ZIP code, birthdate and gender.

AOL released a database of search queries in 2006, with users identified only by a unique number. But the content of the queries quickly allowed individuals to be identified:

> *"..New York Times reporters Michael Barbaro and Tom Zeller, who recognized clues to User 4417749's identity in queries such as 'landscapers in Lilburn, Ga,' several people with the last name Arnold and 'homes sold in Shadow Lake subdivision Gwinnett County Georgia.' They quickly tracked down Thelma Arnold, a 62 year-old widow from Lilburn, Georgia who acknowledged that she had authored the searches, including some mildly embarrassing queries such as 'numb fingers,' '60 single men,' and 'dog that urinates on everything.'"*

This process of re-identification typically requires cross referencing the anonymised data with some other data that includes identities. So one common approach is to make this hard - for example by making the data accessible only on a machine that has no internet connection, and videoing the researcher to make sure they don't take the data away with them. Of course the harder the data is to access, the slower research progress is likely to be.

There is no perfect solution to this issue; privacy projection will never be invulnerable, and comes at a price in terms of clinical progress. In other areas we strike a balance between privacy and practicality, and it seems possible that some mix of de-identification, isolating sensitive data, and penalties for transgressors can be found that will satisfy most people. However, exactly where this balance lies remains a hotly debated issue, and there is no guarantee that the right balance will be found.

5

Conclusion

AI in health is at an earlier stage than the debate sometimes suggests - but its implications could be even more radical.

Current-generation AI seems likely to be adopted in health where there is not much of a competing solution, rather than replacing humans at things they are not good at. Candidates include advice and triage before seeing a doctor, proactive care, and automated second opinions.

On the one hand, AI could significantly enhance health. It could be a force for the democratisation of knowledge and empowerment of citizens, as well as much better health and cheaper healthcare through better prevention, better targeting of treatment and better use of specialist clinical expertise.

But it could also be a disempowering force that reduces citizens' control of their health, creating an expensive and unreliable healthcare system.

Which path is taken depends on choices, so it is important to develop a set of principles to guide the development and implementation of AI in health. We have suggested the following principles and tests:

Principles for People Powered AI	Test
Control. AI should give citizens a clearer and more timely understanding of their health and what should be done, in ways that support greater citizen confidence and control.	Patients should report higher levels of understanding of their condition, control of their health and confidence to manage it.
Simplicity. Well implemented AI should make it quicker and easier for patients to get a resolution to their problem. This requires clarity about the types of problem AI can deal with, and well defined boundaries beyond which human input is required, to avoid AI becoming an additional barrier.	Patients report that it takes less time, fewer steps, and less frustration to get to a resolution of their problem.
Dialogue. The conversation between doctor and patient should remain central. AI should support conversations - ensuring that they are with the right people, that it happens at the right time, and providing the information that supports it. AI should not degrade conversations by over-standardising or taking up unnecessary time.	Patients and professionals should report having higher quality conversations: more time to talk, clearer communication, better mutual understanding and more confidence in the decisions made.
Equity. AI should not be used in ways that exacerbate health inequalities. AI should help all citizens, and most particularly those who face the most challenges and disadvantage in relation to their health and wellbeing.	All previously mentioned metrics, analysed for equity.
Accountability. It must be possible for AI to be understood, questioned and held to account, otherwise AI could fundamentally disempower users - both citizens and health professionals. Without accountability (and the transparency underpinning it), the rest of the People Powered AI principles are hard to achieve - control , simplicity, dialogue and equity all require AI that can be understood and held to account by its users.	Pending European legislation (GDPR) allows for a right to an explanation of a decision from an algorithm. This should be maintained to provide the ability to scrutinise decisions and improve performance.

If the pace of development of AI continues, and given the amount money being invested in it, and the ‘burning platform’ of an overstretched health service AI could become commonplace in as little five years.

There is currently a window of opportunity to shape the future of AI in health. Policymakers are already working to set rules for AI and ownership of public data that ensure the public gets not only value for any data it decides to share, and privacy elsewhere. At this moment they have the leverage to also ensure that AI products that deliver maximum public benefit. This requires that both healthcare providers and users of AI understand the technology, have the tools to shape the market, can understand the needs of citizens, and are able to work through the complexities of implementation. This can be achieved through the following four recommendations:

Public and clinical scrutiny

Involve citizens and clinical professionals in the upstream design, development and implementation of the technology. This should include the requirement of mechanisms, such as public panels made up of citizens, that ensure technology development and implementation takes account of the demands and perspectives of citizens and healthcare professionals, and ensures that People Powered AI principles are applied.

The most reliable guarantors of the principles above are citizens and professionals. It is from their perspective that these issues are most obvious and urgent and the solutions most clear. To gain the benefit of their input, it is critical that they are involved throughout the design, development and implementation process, rather than being brought in late, when key decisions have already been made. Involvement of this kind takes effort to organise, and a degree of innovation in practice. If the public sector is serious about ensuring that AI works for people, this should be a rigorous and mandatory part of the process.

Controlled tests in real-world conditions

Enable real-world experimentation of AI in designated test sites, with non-AI comparators, to understand how AI works in complex systems before wider take-up 'in the wild'.

While general principles can be articulated, we still have a lot to learn about the details of implementing AI and how to design services that take advantage of it while minimising risks. For example, it will require some iteration to ensure safety of advice, while making sure the system is not overrun, or to find the correct balance between use of the AI analytics, and leaving time for human conversations.

This sort of complex technology and service redesign process is hard to do, especially in a responsible way that appropriately manages risks and unintended consequences. Space for rigorous real-world experimentation in controlled areas, before wider roll-out therefore needs to be carved out.

Proactive market design

System leaders actively engage in market design to maximise public benefit and ensure a plural market with genuine choice. This should include regulation that is upstream and proactive ('anticipatory regulation'), clarity over who owns both algorithms and data, and requiring adherence to key design principles, such as People Powered AI principles. Market design should also foster a diversity of new entrants to the market including procurement processes that work for smaller companies and market structures that support a diverse range of R&D activities.

AI development is being driven by private companies, who no matter how well-intentioned, are not directly incentivised to think from the system point of view and bear these principles in mind, or to throttle demand at the cost of a more inaccessable and inhuman system. The public sector needs to shape the market, ensuring it delivers the kind of AI that is good for people and system. This includes, minimally, giving clear requirements, reasonable procurement, and encouraging some disciplined risk-taking. The most effective time to do this is while the general rules for AI are still being set, rather than waiting until providers are well entrenched.

Decision-makers equipped to be informed users

Create a new cadre of public leaders and decision-makers with the technical skills, authority and institutional levers to scrutinise, manage and deploy AI in a responsible way. This should include incorporating artificial intelligence into medical education and health management training to enable the frontline workforce to be informed users of the technology.

Understanding the limits of AI is critical to implementing the technology well, as well as grasping its impact within a complex system. Further, taking advantage of AI will require some reorganisation of services, incentives, perhaps across institutional boundaries, as well as engaging with complex regulatory issues such as data protection. Leadership with the right skills and authority across boundaries is necessary to manage the design and implementation of AI. A health service which has sometimes struggled to get value from its technology investments needs to take special care with an unusually disruptive but high potential technology.

Endnotes

1. Pillay, N., Tisman, A., Kent, T. and Gregson, J. (2010) The economic burden of minor ailments on the National Health Service (NHS) in the UK. 'SelfCare.' 2010;1(3):105-116. http://selfcarejournal.com/wp-content/uploads/2015/09/IMS-1.3.105-16.pdf
2. https://www.pagb.co.uk/content/uploads/2016/06/PAGB_AE_Executive_Summary_June-2015.pdf
3. https://www.babylonhealth.com/
4. Simon, H.A. (1965) 'The Shape of Automation for Men and Management.' New York NY: Harper and Row.
5. Eric Howitz quoted in (2018) 'AI for Good Global Summit Report.'
6. Peng, L.and Gulshan, V. (2016) Development and validation of a deep learning algorithm for detection of Diabetic Retinopathy in Retinal Fundus Photographs. 'JAMA.' 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216. https://research.googleblog.com/2016/11/deep-learning-for-detection-of-diabetic.html
7. Esteva, A. et al., (2017) Dermatologist-level classification of skin cancer with deep neural networks. 'Nature.' Vol. 542, p.p. 115-118; doi: 10.1038/nature21056. https://www.nature.com/articles/nature21056
8. https://skin-analytics.com/
9. parkinsonsvoice.org
10. http://www.bbc.co.uk/news/health-42357257
11. https://alzheimersnewstoday.com/2017/08/30/a-i-big-data-project-predicts-dementia-2-years-before-symptoms-manifest/
12. https://www.theregister.co.uk/2017/07/26/ibm_and_uni_alberta_tackle_schizophrenia/
13. https://diginomica.com/2017/03/17/arthritis-research-uk-enlists-ai-chatbot-arthy-mission-offer-information-advice/
14. https://www-03.ibm.com/press/uk/en/pressrelease/51828.wss
15. Fitzpatrick, K.K., Darcy, A. and Vierhile, M. (2017) Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. 'JMIR Mental Health.' 2017;4(2):19 https://mental.jmir.org/2017/2/e19/
16. BMJ 2017; 358 doi: https://doi.org/10.1136/bmj.j3980
17. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. and Elhadad, N. (2015). 'Intelligible Models for HealthCare. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15.' (pp. 1721–1730). New York, New York, USA: ACM Press. http://doi.org/10.1145/2783258.2788613
18. http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf
19. Christensen, C.M., Grossman, J.H. and Hwang, J. (2013) 'The Innovator's Prescription.' New York, NY: McGraw-Hill Books.
20. https://www.your.md/files/report.pdf

1. Pillay, N., Tisman, A., Kent, T. and Gregson, J. (2010) The economic burden of minor ailments on the National Health Service (NHS) in the UK. 'SelfCare.' 2010;1(3):105-116. http://selfcarejournal.com/wp-content/uploads/2015/09/IMS-1.3.105-16.pdf
2. http://selfcarejournal.com/wp-content/uploads/2015/09/IMS-1.3.105-16.pdf
3. https://www.pagb.co.uk/content/uploads/2016/06/PAGB_AE_Executive_Summary_June-2015.pdf
4. GP at Hand also stands accused of *"cherry picking"* patients - unlike standard GP practices, it discourages some categories of patients as being unsuitable for a largely remote service. As these patients tend to be higher need and higher cost, many critics feel GP at Hand undermines the sustainability of GP services.
5. https://www.digitalhealth.net/2018/01/gp-at-hand-babylon-nhs-england/
6. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

nesta
Health Lab

58 Victoria Embankment
London EC4Y 0DS

+44 (0)20 7438 2500
information@nesta.org.uk
@nesta_uk
www.facebook.com/nesta.uk
www.nesta.org.uk