In brief

Artificial intelligence in healthcare

There are great expectations that artificial intelligence (AI) will help to solve the healthcare challenges of the future, but at the same time, AI brings with it many ethical and societal challenges. AI has been classed as one of the emerging technologies with the greatest beneficial potential but also with the highest risk of negative consequences.[1]

What is AI?

There is no single generally accepted definition of AI. John McCarthy et al., who coined the term artificial intelligence in 1955, defined AI as *the science and engineering of making intelligent machines*.[2] Today AI is often described as computer systems capable of performing tasks that were previously considered to require human intelligence, such as speech and image recognition, translating languages, learning, problem solving and decision-making. The area of research that seeks to develop such systems is also termed AI.

In recent years, developments in AI have accelerated, and every day we read reports of new applications in virtually every area of society. Key factors fuelling the progress made are more powerful computer processors, rapidly growing access to data (Big Data) and the development of increasingly efficient algorithms.

Most current applications of AI are of the type known as *narrow AI*. Here, the purpose is to automate a specific task that is currently performed by humans, usually with the aim of increasing speed, scalability and/or reliability. This can be contrasted with what is termed general AI, where the purpose is to create computer programs that understand and reason at a more general level, like humans do. This publication mainly addresses applications of narrow AI.

Other terms found in discussions of AI are *machine learning* and *deep machine learning* (deep learning). Machine learning uses algorithms that enable an AI system to "learn" to perform a task based on data without having been given specific instructions for that precise task. Deep machine learning is a special form of machine learning that uses algorithms that have been inspired by how the human brain is assumed to

The "In brief" series of publications highlights ethical and societal aspects of a particular topic.

The translation from Swedish has not been reviewed by the Council.

Published May 2020

process impressions and create patterns in order to make decisions.

Another term used is *augmented intelligence*. Here the focus is on applications that support and improve human intellectual capacity rather than replace it. Human and machine become "colleagues" and work side by side. In the healthcare sector it has been predicted that "augmenting" human skills through AI is a more probable development than pure automation, at least in an initial phase.[3]

AI in healthcare

As AI algorithms become increasingly better at sorting data, finding patterns and making predictions, hopes are increasing that AI will be able to take on vital importance in making the healthcare of the future safer and more efficient, while simultaneously helping to solve the challenges we face of an aging population, limited resources and difficulties recruiting qualified staff. Research and development on applications for AI in healthcare is currently being conducted in a large number of fields, from diagnostics and treatment to administration and medical research. However, at the moment there are relatively few systems based on machine learning being used in practical clinical work in Sweden.[4]

Assessments and decisions

There are great hopes that AI will lead to better medical assessments at a lower cost. As AI systems base their predictions on large numbers of patient cases, the hope is that assessments will be able to become more accurate.[5] Unwarranted differences in the practice of individual employees in the health sector might be reduced, along with the risk of mistakes due to cognitive error (information that is ignored, faulty reasoning or bias).[6] AI could also help the healthcare sector to keep guidelines up to date by searching for relevant literature, given that the number of studies in the field of medicine is growing at an ever-faster pace.[7]

Three types of machine learning

In machine learning, algorithms "learn" to solve the task they have been given by being "trained" on large amounts of data. In *supervised machine learning* the training examples contain both input data (e.g. the pixel values in an image) and output data (what the image represents). The algorithm detects patterns in input data that correlate with the output values and creates models to predict the output value when it is presented with new input data.

In *unsupervised machine learning* the training data only comprises input data and the algorithm itself sifts out examples from patterns in the input data. Unsupervised machine learning can lead to the discovery of new, unknown connections.

In *reinforcement learning* an algorithm learns to solve a task by trying different solutions and receiving feedback on how well it has succeeded.

An AI system can be static, which means that the model is not changed once the training phase is over. However, it can also be dynamic, in which the model is continuously adjusted based on new data it encounters.

Diagnostics

A large number of applications of AI have been developed in recent years in image diagnostics. For example, AI has been used to classify skin cancer, find metastases in cancer cases, identify embryos for in vitro fertilisation, detect cervical cancer, assess lung X-rays and diagnose retinal disease at an early stage.[8] Algorithms based on textual analysis have been used to diagnose paediatric diseases.[9] Researchers have also shown that AI-based systems can detect depression based on information from text and speech.[10] In many of the cases, AI has performed as well or better than doctors and considerably faster. In image diagnostics, AI is predicted to take over more even in the immediate future, while the clinical impact in general diagnostics may take longer.[11]

"There are great hopes that AI will lead to better medical assessments at a lower cost."

Prognostics and prevention

AI algorithms have been reported to be able to predict the risk of sepsis (blood poisoning) and cardiovascular disease with higher precision than established assessment instruments.[12] Algorithms have also been developed that broadly seek signs of disease or risk of disease based on the total data found in electronic health records (EHRs): diagnosis codes, prescriptions, lab results, procedures, free text notes, etc.[13] In a study, an algorithm succeeded in predicting, with a high degree of accuracy, long length of stay, mortality and unplanned readmission among hospital patients based on EHR data.[14] AI-based examination of patient data could be used for preventive measures, for example, by raising the alarm if the patient needs to have their drug treatment adjusted or be called in for a check.

Decision-making support system

AI's role in healthcare does not need to be limited to making diagnoses or prognoses. Many people also envisage that in the future AI will come to play an important role in medical decision-making, e.g. by giving treatment recommendations on the basis of diagnoses and other patient data. Some areas in which promising results have been reported are breast cancer and sepsis.[15] Algorithms have also been developed to adapt radiotherapy to the individual with the aim of reducing damage to the surrounding tissue.[16] It has also been proposed that AI might be able to support medical decision-making by predicting a patient's preferences where their decision-making capacity is impaired, based on data in EHRs and social media.[17]

One area where AI is predicted to take on major importance is what is termed precision medicine, which involves tailoring the most optimal treatment for every patient based on genetic factors, medical history, lifestyle, environmental factors, etc. Given the huge amounts of data that need to be analysed, it has been claimed that precision medicine will not be possible without AI.[18]

Other types of decision in healthcare where many see a role for AI are prioritisation decisions, in other words deciding which patient should receive treatment if the resources cannot stretch to everyone, decisions on which patient should receive a donated organ and triaging decisions, i.e. deciding which patient is to receive treatment first and/or at which care level.[19]

Practical healthcare work

It is in assessments and decision-making that the greatest progress has been made in developing AI systems in healthcare. But there are increasing discussions of how AI technology, often linked to robotics, can also be used in practical healthcare work. Examples are AI-supported surgery robots, robots that monitor patients in ICUs, or applications that support telemedicine.[20] The aim can be to increase quality, compensate for staff shortages or, as in telemedicine, bridge physical distances.

Prostheses, etc.

A number of different projects are in progress where AI is used to increase precision and functionality in advanced prostheses.[21] Trials have also been carried out in using AI to control paralysed parts of the body with the brain. In one study, a patient had electrodes that record brain activity implanted in their brain. An algorithm learned to translate this activity into electrical signals that stimulate the muscles in the lower arm.[22]

Patient-focused applications

More and more patient-focused digital applications such as health apps and/or technology worn on or in the body (wearables and internables) are using AI to improve functionality, e.g. by providing more individualised rec-

"Given the huge amounts of data that need to be analysed, it has been claimed that precision medicine will not be possible without AI."

ommendations. AI-based health apps have been developed that assess symptoms and suggest diagnoses or help the patient to find the right level of care in the event of suspected illness.[23] Communication with patients can take place for example via chat bots or computer-generated "virtual health assistants".[24] Other applications have the aim of monitoring and supporting the patients in their self-care, e.g. by helping patients to optimise their treatment or take their medication at the right time and at the right dose.[25] Linked to medical equipment for home use, AI-based apps could be used to monitor patients' health status and sound the alarm in the event of aberrations, which could result in avoiding hospitalisations.

Administration, flows and logistics

Johns Hopkins hospital in the USA uses AI to allocate beds, which has made the hospital's bed assignment process more efficient.[26] Other examples of applications of AI at healthcare administration level are systems for drawing up schedules, uncovering fraud (in insurance-funded healthcare systems), resolving organisational challenges in telemedicine and automating administrative tasks in the health system.[27]

Research and development

There are great hopes that AI could make the development of new drugs more efficient. Algorithms have been developed that use huge amounts of data to learn how molecules interact and so be able to suggest potential drug candidates.[28] In the UK, a database has been built that combines genetic and clinical patient data with multidisciplinary scientific information and uses AI to generate hypotheses for drug development.[29] Other researchers have developed a "robot scientist" that uses AI to make the process of discovering new drugs faster and cheaper.[30]

Additional potential applications in medical research could be finding potential patients for clinical trials.[31] Text analysis algorithms have also proved capable of improving the efficiency of efforts to find relevant studies for scientific review articles.[32]

Ethical and societal aspects of AI in healthcare

Patient safety

Higher patient safety is one of the foremost drivers behind the great interest in AI in healthcare. Many studies have demonstrated positive results in terms of using AI to improve diagnostics, prognoses and medical decision-making. At the same time, there is greater awareness that there are also potential patient safety risks linked to AI.

Atypical cases

An AI algorithm does not learn facts about the world but facts about the database on which it has been trained (see box on p. 5). When it is faced with data and scenarios that deviate from the training data, accuracy can therefore fall. In a clinical setting, this can lead to lower reliability if an algorithm is used on "atypical" patients who have not been represented in the training data, on a patient group for whom the system was not intended or if there are changes to the patient group over time. In one case, all it took was images coming from a different scanner for the reliability to be impaired.[33] Dynamic algorithms that adjust their models as they encounter new cases could counteract such effects but at the same time may lead to less control over the data from which the algorithm learns, which can lead to other undesired effects (see p. 6). The fact that in many cases we do not know how an algorithm reaches its results (see box on p. 9) can make it harder to predict such errors and understand what the errors are due to when they occur.

Differentiating between covariation and causal links

AI algorithms are designed to detect connections in training data that can be used to make predictions when they are presented with new data. In a clinical setting it is vital whether such a link reflects a causal relationship or not. However, today's AI is not capable of making that distinction. An algorithm that was to aid doctors in predicting the risk of dying of pneumonia found that patients with asthma were less likely to do so than the population in general.[34] This contra-intuitive finding is due to the fact that, as a rule, asthmatics with pneumonia are admitted to intensive care immediately and therefore have lower mortality rates. The fact that algorithms do not "understand" causal connections becomes particularly problematic with non-transparent models where it is unclear how the algorithm has reached its results (see box on p. 9).

Skills loss

Even today, much of our collective medical knowledge is stored in medical literature, registers and different health data systems, not with individual practitioners. This trend may be accelerated if AI takes over more and more healthcare tasks and we increasingly trust its decisions. There is a risk that AI systems will take over the role of the repository of collective medical expertise and that there will be a skills loss among staff.[35] This can lead to risks to patient safety were the systems to collapse. Non-transparent systems that healthcare staff do not understand can heighten this risk.

"Automation bias"

If AI support is used for greater numbers of healthcare decisions, there is a risk of what is termed "automation bias", where staff rely too much on automated decision support systems and stop looking for evidence that confirms or contradicts the results. There is a particularly high risk of this with systems that have shown themselves to be generally reliable.[38]

Balancing benefit and risk

The way that algorithms learn means that a system can have high reliability in situations similar to those on which it has been trained, while it can make striking errors if it encounters data that deviate from the training data (see box above). Using AI in healthcare can therefore lead to many patients receiving better and faster care, while some patients are exposed to risks that would not have arisen had they been assessed by a doctor or other member of the healthcare team. One fundamental question is how the risk of errors is to be balanced against the healthcare ben-

Good models demand good quality data

The quality of the data that an AI algorithm uses during training is crucial to the reliability of the algorithm. If the training data contains too few data points or if relevant variables are missing, the algorithm cannot predict output values with sufficient precision. If the examples in the training data instead contain incorrect or irrelevant information ("noise"), the model may be too well adapted to the training data, leading to the system having lower accuracy when faced with new examples. One example is an algorithm for recognising skin cancer which was found to be using the markings that doctors often make on suspected cancer cases as the basis for its diagnosis.[36]

"Artificial stupidity"

During the training phase, an algorithm learns to recognise patterns in the training data. In image diagnostics, the application normally has to go through a large number of images to find patterns that distinguish positive responses from negative ones. However, studies show that what an AI algorithm "recognises" when it identifies a certain type of object can be something completely different from what we humans associate with the same object. This has been found to lead to surprising errors, where things that we would classify as unidentifiable noise are perceived as an object, or where small, insignificant changes lead to the object being completely reinterpreted.[37] Errors of this type could have serious consequences in healthcare.

efits that AI can offer. Is it acceptable that a smaller number of patients risk being harmed if many patients receive better care at the same time?

Medicalisation and over-diagnosis

Increased use of AI-based health apps and wearable technology can lead to early detection of disease and greater independence for the patient. However, constantly monitoring a person's health status could also lead to health stress and to medicalising problems that are fundamentally down to psychosocial factors. Just as with other screening,* algorithms that examine the EHRs of a large number of patients with the aim of detecting disease or risk of disease can result in overdiagnosis and over-treatment and create unnecessary worry in people who are healthy.

Equality and non-discrimination

As with other new methods and technologies that are assumed to be able to lead to improved health for the population AI in healthcare raises the question of how equitable access is to be assured, such that the introduction of AI in healthcare does not risk exacerbating health gaps, e.g. between different socioeconomic groups or between different regions of a country. As with other methods in healthcare, there may be a risk that, for commercial reasons, developers will concentrate on diseases and conditions that many people suffer and that patients with rare diseases will not find it as easy to gain access to the new technology.

Alongside this, due to its particular nature, AI risks bringing about new forms of inequalities or to reinforcing existing threats to equality.

Demand-driven instead of needs-driven healthcare

Higher use of AI-based patient-focused applications can lead to early detection of disease and greater independence for the patient. However, it can also lead to more demand-driven healthcare where patients increasingly seek healthcare for symptoms that a healthcare app has identified, possibly with a desire for a specific treatment that the app has proposed. This could lead to patients with greater needs being crowded out and to inequalities between patients who use such apps and patients who do not. AI-based examination of EHRs to identify risk of disease can benefit patients with high consumption of healthcare and thus a large amount of data in their EHRs, which often reflects need, but not always.

Data-related bias

Many diseases manifest differently in men and women or young people and old people. There are genetic variations between different ethnic groups that may be significant in diagnosis and treatment. Women and older people are often under-represented in clinical trials, while genetic databases often have a preponderance of people of European origin.[39] If a healthcare algorithm learns from a training database in which certain groups of patients are under-represented, it can lead to these groups running a greater risk of misdiagnosis. The reason is that it produces higher accuracy overall if the algorithm adapts its predictions to the largest group.[40] One example is skin cancer, which manifests differently in white

*Tests to detect early disease or risk factors for disease in a large number of symptom-free people.

"One fundamental question is how the risk of errors is to be balanced against the benefits that AI can offer in healthcare."

and dark-skinned patients. An algorithm for diagnosing skin cancer that is mainly trained on white patients therefore risks being less accurate for dark-skinned patients.[41] One question will be whether we are prepared to accept such differences if at the same time it means better diagnosis and treatment for the vast majority of patients.

Even if variables such as age, sex, ethnic origin, etc. do not occur in training data, inequalities or discrimination can arise if variables in input data covary with such characteristics. In healthcare, for example, the use of variables that covary with age could lead to younger patients receiving an unintended advantage in prioritisation situations. This is because in many cases lower age increases the likelihood of a positive outcome.

The choice of output variable, what the algorithm is to predict, can also mean a risk of inequalities in society being reproduced. Studies have shown that American doctors tend to mistake depression for schizophrenia more often in Afro-American patients than in white patients.[42] An algorithm that learns to diagnose disease from previous diagnoses risks reproducing such errors. Another example is an American algorithm used to predict future risk of disease and thus the need for preventive healthcare measures. One study showed that Afro-American patients needed to be in poorer health than white patients to receive the same preventive health initiative.[43] The reason for this was that forecast healthcare costs were chosen as a measurement of future disease risk, something that disfavours Afro-Americans who generally receive less expensive healthcare than white Americans with the same sickness burden.

The fact that many AI systems use non-transparent algorithms (see box on p. 9) can make it harder to detect different types of data-related bias and discrimination and the causes behind it.

Undermining the principle of shared risk

AI-based analyses that derive from patient data could provide increasingly precise predictions of future disease risk. For the patient, this means that preventive steps can be taken. However, such information could also be of interest to insurance companies in order to offer lower or higher premiums depending on the future disease risk of the client. This can undermine the principle of shared risk that is fundamental to all insurance and lead to certain groups in society finding it harder to take out sickness insurance.

Society's assessment of whether a treatment is cost-effective and should be included in the healthcare budget is also based on a form of risk sharing, as both the benefit and the cost usually varies between different patients. With the help of AI, healthcare might be able to make increasingly precise predictions of the individual patient's likelihood of benefitting from a particular treatment. This would thus give rise to the question of whether patients who have lower chances of being helped should also be offered the treatment, or if the cost for these patients would be considered too high in relation to the expected effect.

Built-in values

Assessing different healthcare treatment alternatives demands knowledge of the expected benefit and risk of the different options. However, weighing up expected benefit and risk is not merely a question of facts, but also of *preferences* or *values*. Is, for example, a highly effective treatment with severe side-effects preferable to a somewhat less effective treatment with fewer side-effects? Is the risk of complications from a certain operation balanced by the chance of avoiding life-long dependence on healthcare?

When an algorithm that provides treatment recommendations is designed, it is therefore unavoidable that values and preferences of different kinds will be built in, e.g. through the choice of output variable that the algorithm is to optimise.* The question becomes one of whose preferences and values are built in. Different doctors may make different judgements of which treatment is most appropriate in a given situation.[44] Patients often have different preferences in terms of weighing up expected benefits and risks, for example. The question then is how such differences can be accounted for in an AI-supported healthcare system.

It has also been asserted that if AI algorithms provide concrete recommendations and advice to healthcare staff, they need to incorporate the ethical principles (e.g. beneficence, non-maleficence, respect for autonomy and justice[45]) that we consider should guide our healthcare.[46] However, in order to be applied, ethical principles need to be interpreted and weighed against each other, and the question once more becomes whose interpretations and deliberations are to be built in. Moreover, it is not a given that ethical considerations can always be represented by algorithms, particularly in situations where there is an information deficit and expected benefits and risks are not fully known.[47]

Abuse and malicious use

Unethical design

An algorithm could be deliberately designed to optimise an outcome other than that considered desirable by society. A parallel can be drawn with the Volkswagen scandal where the computer systems of cars were programmed to produce the best outcomes in environmental tests rather than the lowest possible emissions. Instead of optimising the health outcome, an algorithm could be designed to meet set quality criteria without this meaning better care, or to maximise the profits of the healthcare provider, e.g. by proposing measures that are rewarded in the reimbursement system.[48] The lack of transparency of many algorithms can make it harder to detect "unethical design".

Manipulation

The more functions of society are carried out by AI systems, the more incentives there will be to attempt to influence these systems. Because the function of an algorithm is dependent on the data on which it has been trained, one way of influencing a system might be to use cyberattacks to manipulate the algorithm's training database. Users might also be able to attempt to learn how a system works – possibly via other AI algorithms – in order to modify the values that are fed in with the aim of steering the outcome in the desired direction. In healthcare, the manipulation of AI systems could be used, for example, to circumvent treatment guidelines or affect reimbursement (particularly applicable to insurance-based healthcare systems).[49]

Autonomy

As stated, a recommendation from an AI algorithm builds on consciously or unconsciously built-in values. When increasing numbers of large and small decisions are entirely or partly left to algorithms to decide, our choices can be steered in a way that, often without us noticing, risks undermining our power over ourselves, our lives and the society in which we live. This also applies in the healthcare system. If the healthcare offered is steered by algorithms that work out which treatment plan is considered to best benefit the patient – and best meet the patient's preferences – this can lead to restricting patient autonomy. This is particularly true if it is a non-transparent system where the healthcare staff are unable to explain the basis on which the choices were made (see box on p. 9).

*Questions of values are also relevant in diagnostic AI systems. It can, for example, be considered more important that the algorithm does not miss any cases of disease than that it sometimes wrongly decides that a healthy patient is sick. At the same time, too many "false positives" can use up resources that are needed for other patients, which may also need to be factored in.

Using AI to predict future risk of disease through broad analyses of patient data raises questions of what consent should be required from the patient and how the patient's best interests from a health perspective are to be weighed against the right to refuse care and to say no to unwanted health information.

Health apps and wearable technology based on AI can lead to patients having greater control of their own health and greater autonomy. However, constant monitoring of one's own health status can also lead to health stress, where worry about one's own health takes over one's life and restricts autonomy in practical terms.[50]

Privacy

Developing AI algorithms in healthcare takes huge amounts of patient data, including highly sensitive information such as case histories, test results, information reported by patients, diagnoses, etc. This raises general questions linked to Big Data, such as, for example, how companies are able to obtain access to the data from healthcare providers that they need to develop new applications, while protecting the patient's privacy and safeguarding the right to control of one's own personal data.

When it is not necessary to be able to track who the data comes from, e.g. when an algorithm is being trained, the basic approach is that algorithms should only have access to anonymised data. But given sufficient amounts of anonymised data it can be possible – with the help of AI – to reidentify individuals.[51] What are termed synthetic databases are one proposed way of getting round this is. These are constructed based on real personal data such that important information at group level is preserved but each "person" in the database does not represent any actual person.[52]

The black box problem

Many companies see AI algorithms as corporate secrets that are not shared.[53] However, due to their complexity, algorithms for deep machine learning and other advanced AI algorithms are by their nature non-transparent technologies, in which we often do not know how or on what basis an algorithm reaches its results. The lack of transparency of many AI algorithms is termed the "black box problem".

In healthcare, the black box problem creates challenges in terms of patient safety, non-discrimination, autonomy, responsibility and trust. To tackle such challenges, many would like to see greater transparency in AI systems.[54] Some experts have gone so far as to order a halt to the use of all non-transparent algorithms by public agencies.[55] Others consider that the focus should be on developing technologies for explaining how non-transparent AI systems make decisions so as to increase transparency and reduce the risk of error.[56]

One objection is that if we want to exploit the full potential of AI and not slow development – e.g. when it comes to diagnosing diseases with greater precision than humans are capable of – we might have to give up on transparency and the possibility of understanding the grounds on which the decisions are made. The reason is that advanced AI algorithms handle information at a completely different level of complexity than the human brain is capable of.[57] Some even claim that really advanced algorithms are by their nature impossible to explain.[58] What is crucial, however, is not that we understand how the algorithms work – we do not know how all drugs work either – but that they are safe and effective. Therefore, it is said, the focus should not be on transparency and explanations but on validating the reliability of the algorithms in the context in which they are to be used.[59]

Particular problems of privacy are raised in neurotechnology where AI is linked to human brains via brain-computer interfaces (BCI). This kind of technology has already been developed to control prostheses and might become very important in treating diseases such as epilepsy, schizophrenia and paralysis. However, it might also make it possible to decode mental processes and directly manipulate the brain.[60]

Privacy questions can also arise if AI is used for different forms of monitoring, e.g. monitoring patients in the home to prevent hospitalisation.[61]

Responsibility

Healthcare staff perform their work subject to professional liability and can be held legally responsible for incorrect assessments and decisions. Higher use of AI in healthcare raises the question of where the legal responsibility should lie if a diagnosis or a treatment recommendation from an algorithm proves to be wrong. Is it fair to hold the healthcare staff responsible for assessments made by impenetrable "black boxes" whose processes no-one can explain (see box on p. 9)? If not, who should then be held responsible? The programmer? The manufacturer? The agency that approved the system? Or the hospital that used it? Can the system itself be held liable? Perhaps legal responsibility ought to be seen as being shared between several different parties.

To clarify the question of responsibility, it is sometimes asserted that AI in healthcare should work as a "decision assistant" and not a "decision maker" (compare the discussion on augmented intelligence on p. 2).[62] If an algorithm issues a treatment recommendation that does not follow applicable guidelines, the doctor has to assess whether or not it should be followed. Such an approach makes the question of responsibility clear – the decision is the doctor's, who is also responsible for it.*

However, it is not certain that the treatment proposed by the guidelines will be the optimal one or that the doctor's assessment is necessarily better. The algorithm may have discovered that there are cases where the outcome is better with a treatment different from the recommended one.[63] One important purpose of incorporating AI in healthcare is precisely to improve assessments compared with the prevailing situation. If the algorithm is transparent and the doctor understands how it has reached its results, it may be possible to evaluate whether or not an unexpected recommendation should be followed. However, to increase transparency, it may be necessary to accept a reduction in accuracy (see box on p. 9). If we want healthcare staff to take responsibility for decisions founded on AI-based recommendations, we will perhaps have to choose algorithms that are in some cases less accurate.

If AI algorithms become even better and incorrect recommendations increasingly fewer, we may reach a point where doctors are expected to follow a recommendation. It has been claimed that a situation in which a doctor is expected to follow a recommendation from an algorithm, the designer of the algorithm must be held responsible for potential erroneous decisions.[64] One argument in favour of holding developers responsible is that they are in the best position to prevent harmful outcomes. A risk with such a principle is that it may have a chilling effect on the development of new algorithms.[65]

Trust

If AI is to be implemented successfully in healthcare and patients are to be able to share in the benefits that the technology has to offer, patients and healthcare staff alike must be able to trust the AI systems. If the assessments made by the algorithms fail to maintain high reliability, or are less precise for certain patient groups, there is a risk that trust in AI will fall for patients and healthcare staff. Even if the algorithms have high general reliability, trust may be damaged if they sometimes make striking errors that a professional human would not make. This can also damage trust in healthcare in general.

Moreover, patients need to be able to trust that their data is protected. Questions of responsibility must be clear if the healthcare staff are to have the confidence to follow recommendations made by AI algorithms.

*An equivalent situation may arise for other groups of staff, e.g. if a biomedical analyst receives a result from an AI application that deviates significantly from expected.

The question is also how trust is affected if increasing numbers of patients receive diagnosis and treatment suggestions directly from patient-focused AI apps without any human intervention. Will a doctor's professional expertise be valued less highly, and can this reduce trust in healthcare? The question of trust also becomes relevant if the suggestions patients receive from an app differ from those they receive from their doctor, especially if the doctor does not understand how the algorithm works and cannot explain how it reached its suggestions. The question becomes who the patient will trust most.

The black box problem (se box on p. 9) also has a bearing on the question of trust. Transparency and openness with respect to the data used by an algorithm and the logic that controls it can increase trust, partly because it becomes easier to assess reliability. Vice versa, secret or impenetrable algorithms can reduce trust and lead to slowing the implementation of methods that could potentially improve quality. Too much openness, however, seems capable of having the opposite effect on trust and also opens up opportunities for manipulation, which can itself reduce trust.[66]

Today there are many health apps where patients communicate with healthcare staff by video. At the same time, it is becoming increasingly common for patient-focused AI applications to use computer generated "virtual health assistants" in the interface with patients, possibly with the aim of creating trust. When the technology is fine-tuned, these can be increasingly hard to distinguish from real people. This gives rise to the question of how it affects trust in healthcare if it is not certain whether what one is meeting in a health app is a "bot" or a human healthcare provider. Should the patient have the right to always know whether they are interacting with a human or a computer?

Relationships

Greater use of patient-focused AI applications can affect the traditional relationship between the patient and healthcare staff. AI-based health apps and wearable technology can strengthen patients' control of and knowledge of their disease and create a more equal relationship with staff. At the same time, healthcare staff may increasingly come to see patients arriving with ready-made diagnoses and treatment suggestions that they have been given by AI applications. This can lead to a new allocation of roles between patient and healthcare staff where the latter increasingly take on the role of quality assuring or interpreting the information the patient has already received.

As AI systems in healthcare take over increasing numbers of routine tasks, more of the healthcare staff's time could be freed up for patient meetings. In a longer perspective, however, it is possible that a larger proportion of patients' interaction with the health system will be with AI systems instead of people. One question is the impact that not seeing a doctor or nurse has on patient wellbeing and the impact on patient trust in the healthcare system of patients not encountering any people who can convey that trust.

Another consequence, if automated systems increasingly take over more of the tasks hitherto carried out by doctors and other healthcare staff, may be that the traditional relationship between patient and healthcare staff is weakened and the central relationship becomes that between patient and healthcare system.[67] This in turn may lead to undermining the concept of professional responsibility in the health system, which could have negative consequences for patient safety.

"Will the professional expertise of doctors be valued less and can this reduce trust in healthcare?"

Challenges for the future

AI in healthcare has great potential to create benefit for patients and society. AI can play a part in greater accuracy of diagnosis, earlier detection of disease and more personalized treatment. Patient-focused applications can give patients better control of their disease and so increase independence and quality of life. From the perspective of society, AI can help to improve public health and lead to a more efficient use of resources. At the same time, AI in healthcare means several significant challenges of a technical, legal and ethical nature. These challenges need to be highlighted and tackled to prevent the individual from harm and trust in the technology from being impaired leading to fewer opportunities to exploit the opportunities that AI opens up.

QUALITY ASSURANCE. How can algorithms intended to be used in the healthcare system be quality assured and how can it be ensured that the databases on which algorithms are trained and tested are clinically relevant and representative of the patients for whom they will be used? Which processes must be in place for approval and monitoring? By law, Swedish healthcare staff must provide care that complies with science and proven experience, while technical equipment is bound by no such undertaking. This regulatory framework may need to be modified when care is provided to a greater extent by computer systems and not people. Dynamic systems that constantly change their models create special challenges in terms of quality assurance.

TRANSPARENCY. Non-transparent systems can undermine the trust of patients and healthcare staff and make it harder to identify and tackle bias and error. How can the systems be made more transparent? How should the balance between transparency and reliability be determined? Should different degrees of transparency be demanded depending on whether, for example, the issue is one of diagnosis, treatment recommendations or healthcare decisions?

RESPONSIBILITY. Moral responsibility presumes an opportunity and a mandate to exert influence, and sufficient information to evaluate different courses of action. Where does the responsibility lie if an AI algorithm makes an incorrect assessment and this leads to a patient being harmed? How should the legal framework be designed such that responsibility can be demanded at the right level? Should legal responsibility be able to be split between different actors?

EQUALITY AND SOLIDARITY. How do we ensure that the health benefits that AI in healthcare can bring benefit the whole population? How can we ensure that assessments made by AI algorithms in healthcare reflect reasonable values and do not involve discrimination or inequality? How can the principle of shared risk, on which the public health service is based, be maintained if, with the help of AI in combination with large amounts of personal data, we can make increasingly precise assessments of the individual's risk of illness?

AUTONOMY. How are patients' values to be taken into account in AI-supported healthcare such that patient autonomy is not reduced? What rights are patients to have when it comes to having algorithm-based assessments and decisions explained to them so that they can make informed decisions on their care? Should the patient have the right to a second opinion by a doctor? Should there be the right to say no to AI healthcare?

PRIVACY AND DATA PROTECTION. The right to protection of privacy when an increasing amount of data is gathered on an individual citizen is not a question that only concerns AI. Given the large amount of sensitive patient data that is demanded to develop health-related AI, however, it is vital that questions of data protection and the right to control over one's own personal data are resolved, and that a solution is found to the question of how data can be made available to developers in a way that is ethically sustainable.

Conclusion

AI in healthcare that is designed, introduced and monitored in a well thought-out way can bring benefits for patients and society. This presumes awareness of the opportunities and the risks that the technology involves. It is important to take the ethical challenges associated with AI seriously and not to see them as obstacles to innovation but as something that can stimulate and guide development towards applications that foster common goals and values.

The Swedish National Council on Medical Ethics
(S 1985:A)
smer@gov.se
+46 8-405 10 00
@smer_nyheter
www.smer.se

A working group comprising council members Göran Collste, Åsa Gyberg-Karlsson and Nils-Eric Sahlin has worked to produce this report. The main author of the report was research officer Michael Lövtrup.

The Swedish Council om Medical Ethics would like to thank Olle Häggström (Chalmers), Mikael Laaksoharju (Uppsala University) and Stefan Larsson (Lund University) for providing views on a draft of this document.

Further reading

AI4People. (2018). *AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.*

Academy of Royal Medical Colleges. (2019). *Artificial Intelligence in Healthcare.*

AMA Journal of Ethics. (2019;21:E119–197). Artificial Intelligence in Health Care.

European Commission's High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI.*

European Group on Ethics in Science and New Technologies. (2018). *Artificial Intelligence, Robotics and 'Autonomous' Systems.*

Future Advocacy/Wellcome. (2018). *Ethical, Social, and Political Challenges of Artificial Intelligence in Health.*

The Norwegian Board of Technology. (2018). *Artificial Intelligence – Opportunities, Challenges and a Plan for Norway.*

Nuffield Council on Bioethics. (2018). *Artificial intelligence (AI) in healthcare and research* (Bioethics briefing note).

References

1 World Economic Forum. (2017*). The Global Risks Report 2017 12th Edition.*

2 McCarthy J et al. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine,* 27:12–14.

3 Davenport TH and Glover WJ (2018, 29 June). Artificial Intelligence and the Augmentation of Health Care DecisionMaking. *Nejm Catalyst.*

4 Swedish National Board of Health and Welfare. (2019). *Digitala vårdtjänster och artificiell intelligens i hälso- och sjukvården.*

5 Ibid.

6 Davenport TH and Glover WJ (2018, 29 June). Artificial Intelligence and the Augmentation of Health Care DecisionMaking. *Nejm Catalyst*

7 Boggs W. (2017, 14 June). Artificial intelligence may help doctors keep up with new research. *Reuters*

8 Esteva A et al. (2017). Dermatologistlevel classification of skin cancer with deep neural networks. *Nature*, 542:115–118; Liu Y et al. (2019). Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists. *Arch Pathol Lab Med*,143:859–868; Khosravi P et al. (2019). Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ Digit Med,* 2:21; Hu L et al. (2019). An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening. *J Natl Cancer Inst*, 111:923–932; *Axis Imaging News* [website]. (2019, 8 October). AI Detects Key Chest XRay Findings Within 10 Seconds; De Fauw J et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*, 24:1342–1350.

9 Liang H et al. (2019). Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med*, 25:433–438.

10 Hanai TA, Ghassemi MM and Glass JR. (2018). Detecting Depression with Audio/Text Sequence Modeling of Interviews. *Interspeech,* 2018–2522.

11 Obermeyer Z and Emanuel EJ. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med,* 375:1216–1219.

12 Desautels T et al. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inform.* 4:e28.; Weng SF et al. (2017). Can machinelearning improve cardiovascular risk prediction using routine clinical data? *PLoS One,* 12:e0174944.

13 Miotto R et al. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep*, 6:26094.

14 Rajkomar A et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*, 1:18.

15 Somashekhar SP et al. (2018). Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Ann Oncol*, 29:418–423; Komorowski M et al. 2018). The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*, 24:1716–1720.

16 Thompson RF et al. (2018). Artificial intelligence in radiation oncology: A specialtywide disruptive transformation? *Radiother Oncol*, 129:421–426.

17 Lamanna C and Byrne L. (2018). Should Artificial Intelligence Augment Medical Decision Making? The Case for an Autonomy Algorithm. *AMA J Ethics*, 20:E902–910.

18 Mesko B. (2017). The role of artificial intelligence in precision medicine. *Expert Review of Precision Medicine and Drug Development*, 5:239–241

19 *Vetenskap&Hälsa.* (2018, 13 November). Vem ska få nytt hjärta? Artificiell intelligens och läkare gör ofta olika val.; Yaekley RM and Saxena M. (2018, 3 September). How could artificial intelligence benefit emergency medicine? *AI Med* [website].

20 Kassahun Y et al. (2016). Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. *Int J Comput Assist Radiol Surg*, 11:553–568; Smith R. (2019, 7 February). How Robots and AI are Creating the 21stCentury Surgeon. *Robotics Business Review*.; Gholami B, Haddad WM and Bailey JM. 2018, 24 September). AI Could Provide Momentby-Moment Nursing for a Hospital's Sickest Patients. *IEEE spectrum*.; Kuziemsky C et al. (2019). Role of Artificial Intelligence within the Telehealth Domain. *Yearb Med Inform*, 28:35–40.

21 Powell A. (2017, 22 December). AI Is Fueling Smarter Prosthetics Than Ever Before. *Wired.*

22 Bouton CE et al. (2016). Restoring cortical control of functional movement in a human with quadriplegia. *Nature*, 533:247–250.

23 Newton J. (2017, 20 January). Artificial Intelligence App Ada: Your Personal Health Companion. *Medical News Bulletin.; Babylon GP at hand.* (2019). See an NHS GP in minutes for free.

24 Sensely. [n.d.]. *Introducing Ask NHS, powered by Sensely* [website]

25 Twentyman J. (2017, 16 March). Arthritis Research UK enlists AI chatbot 'Arthy' in mission to offer information and advice *Diginomica* [website]; Labowitz D et al. (2017). Using Artificial Intelligence to Reduce the Risk of Non-adherence in Patients on Anticoagulation Therapy. *Stroke*, 48:1416–1419.

26 Javanmardian M and Lingampally A. (2018, 5 November). Can AI Address Health Care's RedTape Problem? *Harvard Business review*.

27 Chun A et al. (2000). *Nurse Rostering at the Hospital Authority of Hong Kong*. Joudaki H et al. (2015). Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature. *Glob J Health Sci*, 7:194–202; Lichtl A. (2018, 11 January). Telemedicine logistics: network optimization using artificial intelligence. *MedCityNews* [website]; Zaidi D. (2018, 22 April). The 3 most valuable applications of AI in health care. *Venturebeat*.

28 Gershgorn D. (2017, 19 April). Artificial intelligence could build new drugs faster than any human team. *Quartz*.

29 AlLazikani B. (2013, 11 November). Artificial intelligence uses biggest disease database to fight cancer. *The Conversation*.

30 Williams K et al. (2015). Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *J R Soc Interface*, 12:20141289.

31 Kent J. (2019, 18 July). Artificial Intelligence Could Increase Clinical Trial Success Rates. *HealthITAnalytics*.

32 Wallace BC et al. (2017). Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. *J Am Med Inform Assoc*, 24:1165–1168.

33 Challen R et al. (2018). Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 28:231–237

34 London AJ. (2019). Artificial Intelligence and Black Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent Rep*, 49:15–21; Caruana, R et al. (2015). Intelligible Models for HealthCare. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15,* doi:10.1145/2783258.2788613.

35 Char DS, Shah NH and Magnus D. (2018). Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *N Engl J Med,* 378: 981–983.

36 Winkler JK et al. (2019). Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatol*, 155:1135–1141.

37 Nguyen A, Yosinski J and Clune J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 427–436; Eykholt K et al. (2017). *Robust Physical-World Attacks on Deep Learning Visual Classification*.; University of Oxford, Department of Computer Science. [n.d.] *Are we safe in self-driving cars?*

38 Challen R et al. (2018). Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 28:231–237

39 Rothwell PM. (2006). Factors That Can Affect the External Validity of Randomised Controlled Trials. *PLoS Clin Trials*, 1:e9; Devlin H. (2018, 8 October). Genetics research 'biased towards studying white Europeans'. *The Guardian*.

40 Zou J and Schiebinger L. (2018). AI can be sexist and racist — it's time to make it fair. *Nature*, 559:324–326.

41 Adamson AS and Smith A, (2018). Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatol*, 154:1247–1248; Lashbrook A. (2018. 16 August). AIDriven Dermatology Could Leave DarkSkinned Patients Behind. *The Atlantic*.

42 Gara MA et al. ((2018). A Naturalistic Study of Racial Disparities in Diagnoses at an Outpatient Behavioral Health Clinic. *Psychiatr Serv*. 70:130–134.

43 Obermeyer Z and Mullainathan S. (2019). Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. I *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)* (pp. 8989). ACM.

44 See, for example Ross C. (2017, 5 September). IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. *STAT*.

45 Beauchamp TL and Childress JF. (2019). *Principles of Biomedical Ethics* (eighth edition). Oxford University Press.

46 Char DS, Shah NH and Magnus D. (2018). Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *N Engl J Med,* 378: 981–983; Dignum V. (2017). Responsible Artificial Intelligence: Designing AI for Human Values. *ICT Discoveries*, Special Issue No. 1.).

47 Sahlin NE. (2018). It's values that matter. I Sahlin NE (red.), *Science and Proven Experience: Johannes* (pp. 77–86). Lund University Publications.

48 Char DS, Shah NH and Magnus D. (2018). Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *N Engl J Med,* 378: 981–983.

49 Finlayson SG et al. (2019). Adversarial attacks on medical machine learning. *Science, 363:1287-1289*.

50 Smer (The Swedish National Council on Medical Ethics). (2017). *Den kvantifierbara människan. Att själv mäta sin hälsa*. Smer report 2017:1.

51 La N et al. (2018). Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning. *JAMA Netw Open. 1:e186040.*

52 Bellovin SM, Dutta PK and Reitinger N. (2019). Privacy and Synthetic Datasets. *Stan Tech L. Rev*, 22.

53 Pasquale F. (2015). *The Black Box Society. The Secret Algorithms That Control Money and Information*, Harvard University Press.

54 See, e.g. the United Nations. (2018). *World Economic and Social Survey 2018*.

55 Campolo A et al. (2017). *AI Now 2017 Report*.

56 Turek M. (n.d.). Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency* (Darpa); Local Interpretable ModelAgnostic Explanations (LIME): An Introduction. *O'Reilly*; Wachter S, Mittelstadt B and Russell C. (2018). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31:841–887.

57 Weinberger D. (2018, 28 January). Don't Make AI Artificially Stupid in the Name of Transparency. *Wired*.

58 Bornstein AM. 2016, 1 September). Is Artificial Intelligence Permanently Inscrutable? *Nautilus*.

59 London AJ. (2019). Artificial Intelligence and Black Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent Rep*, 49:15–21.

60 Yuste R et al. (2017). Four ethical priorities for neurotechnologies and AI. *Nature,* 551:159–163.

61 Smer (The Swedish National Council on Medical Ethics). (2014).*Robotar och övervakning i vården av äldre – etiska aspekter*. Smer report 2014:2.

62 Luxton DD. (2019). Should Watson Be Consulted for a Second Opinion? *AMA J Ethics,* 21:E131–137.

63 See, e.g. Nemati S, Ghassemi MM and Clifford GD. (2016). Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. *Conf Proc IEEE Eng Med Biol Soc,* 2016:2978–2981.

64 Martin K. (2018). Ethical Implications and Accountability of Algorithms. *J Bus Ethics*, doi. org/10.1007/s10551-018-3921-3.

65 Bartlett M. (2019, 5 April). Solving the AI Accountability Gap. *Towards Data Science*.

66 Hosanagar K and Jair V. (2018, 23 July). We Need Transparency in Algorithms, But Too Much Can Backfire. *Harvard Business Review*.

67 Char DS, Shah NH and Magnus D. (2018). Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *N Engl J Med,* 378: 981–983.