


AI4PEOPLE'S 7 AI
GLOBAL FRAMEWORKS



AI IS NOT MERELY
ANOTHER UTILITY THAT NEEDS
TO BE REGULATED
ONLY ONCE IT IS MATURE.

IT IS A POWERFUL FORCE
THAT IS RESHAPING
OUR LIVES, OUR INTERACTIONS,
AND OUR ENVIRONMENTS.

Luciano Floridi

*2018 Chairman, Scientific Committee
AI4People, Professor of Philosophy and
Ethics of Information and Director of the
Digital Ethics Lab at Oxford University.*

AI4PEOPLE'S 7 AI GLOBAL FRAMEWORKS

Following its past work on AI ethics (with the “*AI4People’s Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*”) and on AI governance (with the “*AI4People Report on Good AI Governance: 14 Priority Actions, a S.M.A.R.T. Model of Governance, and a Regulatory Toolbox*”), in 2020 AI4People has identified seven strategic sectors (Automotive, Banking & Finance, Energy, Healthcare, Insurance, Legal Service Industry, Media & Technology) for the deployment of ethical AI, appointing 7 different committees to analyze how can trustworthy AI be implemented in these sectors: the *AI4People’s 7 AI Global Frameworks* are the result of this effort.



HEALTHCARE

Applying a Human-Centered Approach to Assess Risks of Using AI Systems in Healthcare

Authors

Raja Chatila

Chairman Healthcare Committee, AI4People; Professor and Director of the Institute of Intelligent Systems and Robotics (ISIR) at Pierre and Marie Curie University in Paris, France

Stephen Cory Robinson

Senior Lecturer/Assistant Professor in Communication Design at Linköping University, Norrköping, Sweden

Donald Combs

Vice President & Dean of the School of Health Professions, Eastern Virginia Medical School, USA

Paula Boddington

Senior Research Fellow, New College of the Humanities London, UK

Hervé Chneiweiss

Directeur de Recherche au CNRS, Paris, France

Eugenio Guglielmelli

Senior Advisor on Publications for IEEE RAS Professor of Bioengineering Prorector for Research Founder, Research Unit of Biomedical Robotics and Biomicrosystems Università Campus Bio-Medico di Roma

Danny van Roijen

Digital Health Director at COCIR

Jos Dumortier

Honorary Professor of ICT Law at the University of Leuven, Belgium

Leonardo Calini

Policy Manager, European Government Affairs at Microsoft



1. Introduction

On April 8, 2019, the High-Level Expert Group on AI (HLEG-AI) appointed by the European Commission issued the “Ethics Guidelines for Trustworthy AI” (European Commission, 2019a). On June 26, 2019 the group issued the “Policy and Investment Recommendations for Trustworthy AI” (European Commission, 2019b).

In its Ethics Guidelines, the HLEG-AI has identified seven requirements considered key for the design, development, deployment and use of AI systems. AI-based systems complying with these requirements would be considered to be trustworthy and aligned with a human-centered approach.

The HLEG advocated that these requirements become a necessary condition for the adoption of AI systems in Europe.

In its Whitepaper entitled “*On Artificial Intelligence - A European approach to excellence and trust*” released on February 19, 2020, the European Commission summarized its planned AI policy as including “*Policy options for a future EU regulatory framework that would determine the types of legal requirements that would apply to relevant actors, with a particular focus on high-risk applications.*” (European Commission, 2020a)

The risk-based approach defined in the Whitepaper is actually based on a two-tier definition of risk.

To be considered “high-risk”, the AI system must be deployed in a sector known to be high-risk, e.g., the healthcare sector. Second the AI system must be used within this sector in an application, which is itself considered high-risk.

A subsequent report by the HLEG-AI, “*The Assessment List for Trustworthy Artificial Intelligence*” (ALTAI), was published on July 17, 2020 in an effort to provide an initial, more concrete, approach to evaluating compliance with the Ethics Guidelines for Trustworthy AI systems (European Commission, 2020b). The HLEG-AI stated that:

“The Assessment List for Trustworthy AI (ALTAI) is intended for flexible use: organisations can draw on elements relevant to the particular AI system from this Assessment List for Trustworthy AI (ALTAI) or add elements to it as they see fit, taking into consideration the sector they operate in. It helps organisations understand what Trustworthy AI is, in particular what risks an AI system might generate, and how to minimise those risks while maximising the benefit of AI. It is intended to help organisations identify how proposed AI systems might generate risks, and to identify whether and what kind of active measures may need to be taken to avoid and minimise those risks. Organisations will derive the most value from this Assessment List (ALTAI) by active engagement with the questions it raises, which are aimed at



encouraging thoughtful reflection to provoke appropriate action and nurture an organisational culture committed to developing and maintaining Trustworthy AI systems. It raises awareness of the potential impact of AI on society, the environment, consumers, workers and citizens (in particular children and people belonging to marginalised groups). It encourages the involvement of all relevant stakeholders. It helps to gain insight on whether meaningful and appropriate solutions or processes to accomplish adherence to the seven requirements are already in place or need to be put in place.”

The combination of the Guidelines and the Assessment List provide a solid foundation for the assessment of high-risk AI systems operating in high-risk sectors such as healthcare.

Following its past work on ethics and governance for AI (Floridi, 2018; Pagallo, 2019), AI4People has identified healthcare as one of the strategic sectors for the deployment of AI, and has appointed a working group to analyze how trustworthy AI can be implemented in this sector, which is considered high-risk. This paper examines how the seven requirements are relevant and can be used, along with other tools such as ALTAI, to assess risk in the deployment of AI systems in healthcare.

The aim is to illustrate a practical approach to assessing risk and to provide recommendations to stakeholders in this sector. It is important to note here that risk in healthcare, and thus in using AI in healthcare, is multi-dimensional. This multi-dimensionality will be explored through two examples of use cases that illustrate how risk can be assessed.

2. AI and healthcare

Healthcare is complicated. It involves assessing the current and trending state of health among patients with differing genetic makeups, personal history, environmental exposure, behavioral patterns, social contexts, cultures, economic status, self-awareness and patterns of healthcare usage.

Healthcare practitioners use various kinds of data for decision-making, including diagnostic laboratory and radiologic tests, written notes and electronic health records recounting patient interviews and anamnesis, data about the history of family health conditions, epidemiologic models of infectious diseases, and knowledge of available resources. They often use these data in situations of high urgency. The amount of data available from all these sources overwhelms the processing capacity of practitioners. It is therefore no surprise that AI systems find a great number of applications in this domain, where they promise faster and more comprehensive decision-making support



than practitioners can muster on their own. One substantial clinical application of AI has been in the imaging professions (Ting, 2018) - radiology and sonography - where, thanks to data availability and to progress in the development of effective algorithms, AI systems have shown a high level of accuracy, helping to identify tumors in breast cancer, retinal disease and recently even fast diagnosis of COVID-19 pulmonary diseases (McCall, 2020).

Recognizing the progress that has been made does not mean, however, that the AI systems should be already considered fully trustworthy. Generally speaking, many algorithms based on deep learning techniques are considered black boxes (Castelvecchi, 2016; Barredo Arrieta, 2020). Regarding interactions of practitioners and patients with AI systems, there is little understanding of the degree to which practitioners defer to the algorithms. And, finally, the big question of who is ultimately responsible for the diagnosis, treatment and outcomes of healthcare has not been answered yet.

3. Risk, danger and hazard

The notion of **risk** used in the EU Whitepaper (European Commission, 2020a) and in the ALTAI (European Commission, 2020b) needs to be clarified in order to analyze how to appropriately assess AI-based systems in healthcare.

A **risk** is a possible harm, more or less foreseeable, measurable by a probability of seeing a danger materialize, while the hazard is an unpredictable and unexpected event, even if it could be probabilistically modelled.

A **danger** is the presence of a factor that compromises the integrity, security, wellbeing, or existence of a person, an entity or an object. A danger may remain without risk, if one knows how to avoid it completely, while a risk always has at its source a danger, which must be identified. For example, in healthcare, a danger might be an infectious pathogen and the risk is the frequency with which an individual develops the corresponding disease.

A **putative risk**, not grounded in scientific or empirical evidence, must also be distinguished from a proven risk. Indeed, a proven risk is never zero. For example, although air travel is the safest form of transportation, the risk of a plane crash is not zero (the probability is estimated to be about 10^{-7} for current airliners). On the other hand, a putative risk can become zero. For example, in 1836 François Arago, famous scientist and mathematician asked authorities to prohibit people from riding trains because he foresaw a major danger for health beyond the speed of 27 km/h and while traversing tunnels (Arago, 1836). He believed, incorrectly, that the human body would not resist the pressures produced.



The introduction of new technologies in our society can generate not only benefits, but also dangers and risks which must be properly assessed and managed. This is frequently the case, even for very relevant and popular technologies. For instance, automotive technologies are widely diffused and appreciated, but, according to the World Health Organization (WHO, 2020), they are the first cause of death for citizens aged 5-29 years and they have an economic impact, mainly in terms of cost for the healthcare systems, of 3% of the gross domestic product worldwide. Ensuring a safer system approach for all road users is one of the main goals of the UN 2030 agenda for Sustainable Development, which requires important innovation on automotive technologies, including AI-based solutions forgiving human errors. Generally speaking, **governing the introduction of new technologies** in our society clearly requires also to elaborate and promote guidelines, policies and regulatory issues to prevent and mitigate potential risks. These examples call for a precautionary approach to evaluate the reality of dangers, and an evaluation of the resulting risk so that unnecessary measures are not taken for nonexistent or very low risks, and appropriate measures are taken to mitigate proven risks.

When we consider AI-based systems in healthcare, the stakes are high because we are dealing with human life. There are potential dangers, for example, an interpretation mistake in medical imagery could lead to a cancerous tumor going unnoticed; a mishandling or lack of security measures for health data could lead to the disclosure of patient personal data. It is therefore important to correctly qualify the actual dangers of specific technical solutions and to accurately evaluate the related risks so that AI systems are deployed for the benefit of patients and society.

The question is: how to define risk indicators that make it possible to identify if there is a risk at all. For example we could ask the “worst case scenario” questions: *Are there catastrophic consequences of a system failure?* And *“To what extent could the system be considered dependable, i.e. capable of mitigating associated risks for users?”* By performing such analyses, the occurrence probabilities of those events leading to system failure have to be considered, and appropriate measures taken to reduce them to safely manage failure implications and prevent failure repetition. This precautionary process can imply *e.g.*, system redesign, different use protocols, clearer interfaces, user training, and assessment of user capabilities. These measures are classical in critical applications.

The notion of “high” vs. “low” risk underlies a kind of threshold, under which risk could become acceptable. However, using such a binary scale (high/low) might be too limited to express risk impact diversity. A risk scale expressing damage intensity should be *multidimensional*, accounting for different values that could be at risk. For example, data privacy, physical integrity, physical wellbeing, moral impact. In each dimension, risk could be evaluated taking into account several parameters such as



patient context, *e.g.*, age, lifetime expectancy; medical history; healthcare general context, *e.g.*, availability of means or equipment, or of alternative treatments; impact on the healthcare system itself. Risk should also be considered over time, to assess short term and longer term impacts.

Finally, the scale should be a continuum to avoid arbitrary and *a priori* thresholds. This implies a degree of complexity that would be difficult to capture with a binary “high/low” scale.

Dimensions of risk Influencing factors	Physical integrity	Physical wellbeing	Mental wellbeing	Privacy & intimacy	Agency & autonomy
Personal context					
Personal health condition					
Personal health history					
Age and lifetime expectancy					
Care means availability					

Figure 1. Example of dimensions of risk (or at risk) and factors that influence them. The time dimension is also to be considered due to cumulative effects

The danger/risk analysis assessment and risk mitigation should be performed from the onset of system specification, and continually, after deployment throughout the system’s life cycle, taking into account standards and certification processes.

A sound methodology should be developed to correctly make these evaluations and to mitigate risk. In the domain of software engineering, solid concepts and methodologies have been proposed to deal with the dependability or resilience of software systems. Dependability (Avizienis, 2004), defined as “*Delivery of service that can justifiably be trusted*” has several attributes, including system availability (readiness for correct service), reliability (continuity of correct service) and safety (absence of catastrophic consequences on the user(s) and the environment). “Justifiably” means that there is a grounded and proven assessment of these properties. The notion of danger underlies catastrophic failure consequences. Limiting consequences of task failure includes verification and validation techniques, such as error detection and recovery mechanisms, model checking, detection of incorrect or incomplete system knowledge, and resilience to unexpected changes due to environment or system



dynamics. There are means to reach these objectives, such as software system design diversity, redundancy, as well as software architectures enabling system state assessment for decision-making in order to produce error-free results.

This last step may however be performed by a human specialist for example, and requires a specific protocol. For example, the ALTAI (European Commission, 2020b) could be a guide here. This raises issues related to the organization and governance of the healthcare system, and not merely of a piece of software providing a given service. The combination of risk assessment and decision-making is actually a source of complexity, because there is a cost in making the systems fail-safe. Eliminating dangers, i.e., reducing risks while keeping benefits might indeed incur important investments in time and finances as we can learn from the aviation industry for example - hence the 10^{-7} probability of an airliner crash. But we can also see this approach in healthcare, especially in the pharmaceutical industry and for the design of medical devices. AI systems add a level of difficulty when they are based on learning methods, which are opaque and as such challenge classical verification and validation techniques. A whole field of research currently addresses transparency and explainability issues of AI systems (Barredo Arrieta, 2020). Also, health authorities have already issued guidelines for assessing medical devices which include AI-based systems, *e.g.*, in France (Higher Health Authority, 2020).

4. Case studies

In order to discuss a risk-based approach, we analyze next two case studies that illustrate the use of AI systems in healthcare in order to identify where the implementation of AI systems might bring potential benefits (improvements to treatment outcomes and diagnostic accuracy, healthcare system efficiency, etc.) and to highlight ethical issues, specifically the seven key requirements for trustworthy AI identified by the HLEG-AI (which are the basis for the ALTAI), when implementing AI technologies in the healthcare sector considering a risk evaluation approach as defined in the EU Whitepaper.

1. AI systems for patient triage and prioritization, also dealing with crisis situations such as the COVID pandemic;
2. AI systems for diagnosis.

Case 1. Patient triage and prioritization

When the waiting list of patients is quite long, and the diversity and urgency of healthcare that is sought is multifaceted, an AI system can help to compensate for the lack of adequate personnel to deal with the flow of patients. Recommendations from



an AI healthcare system can help to thoroughly analyze patients' healthcare records in combination with their presenting symptoms.

Based on these factors, the healthcare staff will be able to prioritise and treat those with the most urgent needs. Note also that emotion detection could be used in such systems, which would raise additional ethical issues which are beyond the scope of this paper (see (Grandjean, 2008; Greene, 2019)).

Patient triage and prioritization can be done through an AI system interacting directly with the patients, with the healthcare personnel, or with both. Question and answering systems, or chatbots, which are likely to be the interface to the AI system, will sort the patients to the appropriate level of urgency through a dialogue. The challenge is whether the AI embedded within, or connected to, the chatbot will reliably triage patients to the appropriate level of care. Triage involves a combination of complicating factors--the communication skills of both the AI system and the patient, and the assumption of a factual description of the current symptoms and relevant physical and mental history to mention only two such factors.

A “chatbot” or Artificial Conversational Agent (ACA) is a software system that has natural language processing (NLP) capacities enabling it to enter in a dialogue with a user through a keyboard or a voice recognition and synthesis systems and could also use a visual avatar. One of the first such systems was ELIZA developed by Joseph Weizenbaum at MIT in the mid-1960s, which was based on using keywords and scripts. Interestingly, ELIZA's scripts were based on reformulating user inputs as questions to her/him in a way resembling the communication strategies of Rogerian psychotherapists (Weizenbaum, 1966).

Some of the most known and popular chatbots today are commercial systems such as Amazon's Alexa, Google Home, or Apple's SIRI, connected to the Internet and thus able to access considerable data to answer questions or to conduct e-commerce. Chatbots are also integrated in several specific systems, such as GPS car route planning or queries for travel companies on their websites. Some systems, such as those mentioned above, include a learning capability, see (Kim, 2018), enabling them to improve their response according to new data, previous choices made by the user, or exploiting inputs from other users.

General ethical issues with Chatbots

There are several ethical issues related to developing and using chatbots, and a few of them can be exacerbated when healthcare becomes the application domain. To list but a few:



- *The users might not be aware that they are actually interacting with a computer program and not a human being.*
- *The chatbot's voice and/or appearance might have a specific tone or aspect that might influence user behavior.*
- *The chatbot's behavior will be based on algorithms, which might include AI and learning capabilities, and on a variety of data. Similarly to all such systems, the data may be biased and the chatbot language or behavior as well.*
- *Like all algorithms, including AI systems, a chatbot lacks semantics and does not actually understand what it is doing or what consequences its outputs might have on humans.*

Risk assessment for chatbots in healthcare

Chatbots are also used in healthcare, e.g., in psychiatry (Philip, 2020). In a healthcare context, one must distinguish the “operator” i.e., the medical professional (or organization), which deploys the chatbot, from the “user”, the person who is going to actually interact with it through Q&A, from the device manufacturer. The operator is generally not aware about the internal workings of the system, but knows how to use it and what to expect from it. The user is often totally ignorant of the underlying algorithms of the chatbot and its capacities. There are different issues to consider from these two perspectives.

The consequences of chatbot advice or decisions might be severe for the user. One main issue is related to the fact that the chatbot ignores the general context in which it is used, and can only use specific information about patient condition and possibly their medical data. This does not mean the decisions can be wrong. On the contrary, sometimes, and often, the decisions are correct and suggest that the system has been well designed and trained. However, the risk related to wrong decisions remains high because of the consequences for the health of the patients. This has to be acknowledged as a factor for deploying chatbots and evaluating their conclusions by the operator.

Furthermore, correct decisions will tend to increase operators' confidence and trust in the system, perhaps leading them to not question the triage decisions over time.

The chatbot might influence the user through the form and content of its questions and answers, thereby inducing a bias in the user's behavior, that may, in turn, produce a bias in the chatbot decisions. In the instance of an incorrect triage decision, that could prove catastrophic.

The chatbot might not be able to say “I don't know” unless it's explicitly programmed to do so, and might persist in forcing the dialogue to acquire additional data, orienting users' answers. This might produce inappropriate concluding decisions.



Trustworthy AI elements

Almost all of the seven requirements for “Trustworthy AI” need to be considered in evaluating the chatbot AI system, given the ethical questions raised above.

1. Agency: Patients (users) should be informed that they are dealing with a machine (see transparency) and should have the possibility to opt-out and to access a human. Operators should be able to assess and validate the results of the chatbot decisions through metrics (e.g., confidence, performance, explainability).
2. Technical robustness and safety: Chatbots should be verified and validated by certification bodies or trusted third parties.
3. Privacy and data governance: Data collected by the chatbots and underlying platform should respect and comply with general regulations as well as health data sensitivity (anonymity, proportionality, purpose of use, storage and access) as recognized by the GDPR (EU 2016/679).
4. Transparency: The patients should be clearly informed that they are conversing with a chatbot and not with a human through an interface. The purpose of using a chatbot should be stated. The professional operator should be informed about the system’s decision process
5. Diversity: The chatbot interface (voice, visual appearance, attitude) should be as neutral as possible and its makers should not try to give the image of a human in its visual appearance to avoid confusion. Issues of diversity are relevant, specifically where facial recognition and emotion detection are utilized for patient sorting. As documented widely, certain ethnic groups can suffer erroneous facial recognition detection, such as populations of color (Grother, 2019). These false-positives might include in addition incorrect emotion detection, which jeopardizes the entire concept of fair patient sorting based on real-time behavioral responses including emotion, etc.
6. Accountability: Accountability and liability must remain with human beings (designers, operators, users, etc.) and not on the machine itself. Indeed, AI systems should not have a legal personality.

It is possible to examine the risk assessment process more thoroughly by applying the ALTAI to one of the trustworthy requirements. The fourth guideline for trustworthy AI addresses transparency, which encompasses three elements: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system. In developing an approach to assessing risk, the HLEG posed some illustrative questions that operators and users of AI in healthcare might employ to identify and mitigate risk (European Commission, 2020b). Their discussion is worth excerpting in the next few paragraphs.

Traceability



This subsection helps to self-assess whether the processes of the development of the AI system, i.e. the data and processes that yield the AI system's decisions, is properly documented to allow for traceability, increase transparency and, ultimately, build trust in AI in society.

- Did you put in place measures that address the traceability of the AI system during its entire lifecycle?
 - ◊ Did you put in place measures to continuously assess the quality of the input data to the AI system?
 - ◊ Can you trace back which data was used by the AI system to make certain decision(s) or recommendation(s)?
 - ◊ Can you trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system?
 - ◊ Did you put in place measures to continuously assess the quality of the output(s) of the AI system?
 - ◊ Did you put adequate logging practices in place to record the decision(s) or recommendation(s) of the AI system?

This could take the form of a standard automated quality assessment of data input: quantifying missing values and gaps in the data; exploring breaks in the data supply; detecting when data is insufficient for a task; detecting when the input data is erroneous, incorrect, inaccurate or mismatched in format – e.g., a sensor is not working properly or health records are not recorded properly. A concrete example is sensor calibration: the process which aims to check and ultimately improve sensor performance by removing missing or otherwise inaccurate values (called structural errors) in sensor outputs. This could take the form of a standard automated quality assessment of AI output: e.g., predicted scores are within expected ranges; anomalies in output are detected and input data leading to the anomaly detected and corrected.

Explainability

Assessing the explainability of the AI system is a second element of trustworthiness. This element refers to the ability to explain both the technical processes of the AI system and the reasoning behind the decisions or predictions that the AI system makes. Explainability is crucial for building and maintaining users' trust in AI systems. AI driven decisions – to the extent possible – must be explained to and understood by those directly and indirectly affected, in order to allow for contesting of such decisions. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contribute that) is not always possible. These cases are referred to as 'black boxes' (Castelvecchi, 2016) and require special attention. In those circumstances, other explainability measures (e.g. traceability, auditability and transparent communication on the AI system's capabilities) may be required, provided



that the AI system as a whole respects fundamental rights. The degree to which explainability is needed - which depends on whom it is intended to (Barredo Arrieta, 2020) - depends on the context and the severity of the consequences of erroneous or otherwise inaccurate output to human life (European Commission, 2020a).

- Did you explain the decision(s) of the AI system to the users?
- Do you continuously survey the users if they understand the decision(s) of the AI system?

Communication

This subsection helps to self-assess whether the AI system's capabilities and limitations have been communicated to the users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy as well as its limitations.

- In cases of interactive AI systems (e.g., chatbots, robo-lawyers), do you communicate to users that they are interacting with an AI system instead of a human?
- Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?
 - ◊ o Did you communicate the benefits of the AI system to users?
 - ◊ o Did you communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/ or error rates?
 - ◊ o Did you provide appropriate training material and disclaimers to users on how to
 - ◊ adequately use the AI system?

Case 2. AI prediction (...and outcome/diagnosis reassessment)

AI has the possibility for greatly changing healthcare – from cost savings through more efficient healthcare, preventing physician burnout by lessening administrative tasks and increasing direct patient care - the use of artificial intelligence in the healthcare environment will be vast and the potential improvements immense. Specifically, the ability for AI to 1) predict onset of health conditions leading to proactive healthcare interventions, or 2) through reaffirming or rejecting physician diagnoses, or finally 3) assisting in patient healthcare record management, can ultimately result in saving lives or increased patient quality of life.

One scenario where AI can improve healthcare outcomes is by predicting onset of health conditions/diagnoses, leading to capabilities to proactively manage healthcare. For example, if a patient is genetically predisposed to cancer, AI can be utilized in personalized medicine to flag the patient's risk of cancer, and recommend healthcare



interventions for the patient. These scenarios are not hypothetical, but could become reality in Denmark and Estonia. In Estonia, for example, there already exists precision prevention for breast cancer or cardiovascular diseases, this is enabled by the nation's biobank (Milani et al, 2015), which currently houses genetic information for 5% of the population. The use of AI could increase the speed of identifying citizens/patients facing these potential diagnoses. Use of AI in routine patient diagnosis could also assist in basic research into disease including classification, prognosis, and treatment.

A second scenario where AI can benefit both healthcare providers and patients is the ability to reconfirm or "correct" healthcare findings (i.e., tests, diagnoses) - similar to the partnership between the UK's NHS and Google's Deepmind AI (King, 2019), healthcare institutions might utilize AI to verify physician findings. In scenarios of telemedicine, such as video meetings, the AI might verify a small sample of digital healthcare meetings to reaffirm that telemedicine is achieving similar, or on-par, results found in face-to-face healthcare.

A third scenario entails AI assisting in patient health record management. For example, after a patient visits their physician, the resulting notes of the visit can be automatically transcribed and added to the patient's health records. The AI can also determine the next appropriate steps for the patient, as well as schedule appropriate follow-up visit communications. The AI system provides advice about the patient condition and further actions to be taken by the physician. In essence, the AI assists the physician by managing the healthcare record.

Given all these benefits from the adoption and implementation of AI in healthcare, there also come corresponding risks and consequences arising from AI in healthcare.

Risks analysis for AI predictions and outcome assessment in healthcare

Various types of risk exist.

Some risks pertain to the quality of the data. AI could potentially both predict and reaffirm health diagnoses. However, some uses of AI to reach medical diagnoses and recommendations may be flawed; there will be a need for ongoing research and transparency (Antun, 2020).

Issues related to poor, erroneous, or incomplete patient data are significant when AI is utilized for maintaining patient records. Training data may be poor, for example, failing to adequately reflect patient groups, may not adequately reflect variations in record keeping (Panch et al., 2019).



Healthcare meetings are complex human interactions and an AI system is likely to miss many elements of human importance. Subtle biases may be incorporated in data recorded in health records which the AI may a) fail to notice and hence reproduce or b) be able to detect and address (Char et al., 2018). Patients' medical data might be located in different electronic health record systems (EHRs enable patient data to be digitally accessed in one central file, allowing stakeholders to seamlessly share, access and exchange a patient's health information (Shortliffe, 1999)), or include data from wearables and other sensors - AI might pose a solution for interpreting data from different sources and different classifications of data, however AI-based solutions are early and mainly used in medical research (Lotman & Viigimaa, 2020).

Some risks may arise from the nature of the data required. Verification of results could be based on a variety of indicators including medical outcomes but also patient and physician satisfaction.

Assessing success of such healthcare meetings thus may necessitate facial recognition and emotion detection AI. This could have risks of discrimination and labelling of certain patients. Situations posing certain risks may arise from the combination of human and AI expertise.

Suppose a physician disagrees with all or part of the AI's recommendations - a system may not allow this; conversely a well-functioning system may be overridden. An institution should develop protocols to deal with such situations.

Mitigation of many risks includes attention to issues beyond AI itself, both within and beyond the medical setting. Within a medical setting, some of these concern pre-conditions for the successful use of AI, some concern possible longer-term impacts of its use. For the AI to function effectively in managing patient health records and advising follow-up communication, prior work is needed integrating computing systems across different sectors. Without this, gains may be fragmentary and illusory (Panch et al., 2019).

There is a possible risk of impact on developing physician's skills and learning from clinical experience, which would need to be monitored and addressed. This is also necessary for good communication from physician to patient regarding their condition and recommendations.

There is a risk of focus on certain technology such as AI at the expense of necessary work on other technologies and the importance of the clinics (symptoms and real-life experiences of the patient).

Faster and easier detection of very early disease stages and focus on risk carries benefits but its routine and long-term use also complex questions pertaining to issues such as risk perception and medicalization which may require relevant expertise to address (Featherstone et al., 2020).



Wider economic and legal issues may arise. Possible flagging/screening by insurance companies of an individual's terminal illness before onset, may lead to exclusion of these patients from the healthcare insurance market. Here, the predictive screening capabilities of AI will not result in saving lives or better healthcare outcomes, but in creating patient discrimination and possible surveillance of those with sensitive or undesirable healthcare conditions/disease/diagnoses (HIV, covid19, etc).

A general unknown risk concerns the future of litigation and case law in medical practice from 'bad cases'. What will happen to the current relationship between a patient and individual physicians with the use of AI systems in recommending treatment? (Char et al., 2018).

Trustworthy AI elements

1. Human agency: such use of AI should not unnecessarily override individual medical judgement and autonomy. Protocols for dealing with mismatch between the judgements of physician and of AI systems raise the risk that this may not always function in the best interests of each individual patient, may for instance be guided by fear of litigation, and/or by focus on certain audited risks rather than on other less tangible risks which are not audited. Will there be room for genuine difference of medical opinion (Char et al., 2018)? Further, patient and physician AI education will be needed for humans to fully comprehend the impact of AI in healthcare – without "AI literacy" humans are not able to fully embrace and protect their own agency.
2. Technical robustness and safety: as outlined in the last section, reliance on AI must not be premature. The AI systems deployed in medical care must be both technically robust, and their technical safety ensured through repeated auditing of such systems (Raji et al., 2020). Further, one of the biggest problems facing citizens is the lack of information about the types of data analyzed in AI systems (Vinuesa et al., 2020). Both issues can be partially rectified by mirroring the use public registers of algorithms used in Helsinki and Amsterdam allowing auditing of such AI systems (Johnson, 2020), allowing citizens to identify the databases that trained the model, how individuals utilize the prediction, description of how each algorithm is used, and how bias or risks were assessed in the algorithms.
3. Privacy and data governance: The protection of sensitive health data utilized in AI-assisted healthcare is not only powerful in its ability to deliver targeted, personalized healthcare, but also has significant issues for potential discrimination or surveillance of patients. Data should remain subjected to GDPR rules which should be strictly applied. Healthcare institutions are based



on trust – trust in the physician, trust in the healthcare institution, and trust in the sanctity of patient data. When trust is broken (not an “if”, but a “when”), it is key to identify whom’s data was breached, which specific data (i.e. diagnoses or prescriptions), and subsequent potential for fraud or discrimination must be minimized with haste. Critically, individuals should be required to provide clear, meaningful consent for use of their data in healthcare making decisions powered by AI systems, which would enable a better data traceability.

4. **Transparency:** Transparency and explanation to the physician will be needed at a high level. The importance of checking may be highest for patient groups with reduced capacity to understand the involvement of an AI system, such as those with cognitive impairment.
5. **Diversity:** imposing uniformity on health care records may be counter to nuances needed to accommodate different groups. Conversely greater ease of personalized medicine and diagnosis may assist in fine-tuning diagnosis and treatment for groups whose disease presentation and treatment may differ from the average of the population. Additionally, it benefits all healthcare institutions to ensure that diverse training sets are utilized in order to make public health decisions. In facial recognition systems, we have seen a lack of diversity in training sets where the algorithms resulted in poor identification of individuals of color (Maurer, 2017, Merler et al., 2017) and therefore databases used for training must hold diversity (in the data) as sacrosanct.
6. **Societal and environmental wellbeing:** increased diagnosis and medicalization can have downsides as well as benefits, including increasing healthcare costs, weighed against increases in preventative health and personalized medicine which may save costs both monetary and personal costs to the patient of unnecessary or delayed treatment. Unknown risks relate to the possible impact on litigation with complex questions for medical professionals, patients, and society as a whole, including risks of increasing litigiousness in medicine. Societal wellbeing can be jeopardized (including public trust) if debacles such as the NHS’ “care.data” scandal are not learned from (Vezyridis & Timmons, 2017). Individuals not able to practice informed consent must be protected and prioritized, as well, as AI brings with it many issues of comprehension and public understanding.
7. **Accountability:** There will be a certain amount that is unknown about how the law might develop in this area so hospital managers and those in charge will have a responsibility to monitor such situations carefully. Individual medical practitioners and patients also need protection and caution as the full implications become apparent.



5. General discussion and conclusions

The amount of information - from databases of increasing diversity, from the proliferation of sensors, and from smartphone-based apps linked to electronic health records, just to mention a few drivers of change - is beyond the capacity of human intellectual processing. For that reason alone, there will be a steadily increasing use of AI applications by medical and health professionals and the organizations in which they work. These applications will improve healthcare, but they also have the potential to introduce new risks from AI for both patients and professionals. However, trust in purpose and in operation is the foundation for the development and adoption of technologies, and AI is no exception.

Given that healthcare is a high risk sector, these additional sources of risk are beginning to be addressed through the development of requirements for trustworthy AI and tools for assessing risk such as the ALTAI. This report has examined the issue of risk and approaches to assessing and managing risk in healthcare. We have illustrated an approach or use case as to how the seven elements of trustworthy AI might be merged with the ALTAI lists developed by the HLEG to assess some of the risks associated with the use of chatbots in a healthcare setting. The primary argument is that asking focused questions about AI in a specific application or setting from the perspectives of operators and users can help to determine risk and trustworthiness. Our purpose is less to provide specific guidance for assessing trustworthiness of AI applications and more to suggest that developers and users need to take responsibility for developing an appropriate assessment process in their particular setting.

In summary, the following findings have been ascertained:

- Complying with a human-centered AI approach can be assessed through the compliance with the 7 key requirements.
- Risk is not binary. There is a multidimensionality in its nature, a continuum in its intensity as well as a time factor. Assessing risk requires identifying the values that are impacted and the degree to which they are.
- Healthcare is by nature a domain of high stakes. It is also a domain in which several factors are interrelated. A hospital procurement policy may impact its ability to cope with emergency situations. Its management of appointments may impact the availability of beds or operation rooms. It is difficult to assign a priori a risk level to such or such application.

Another important issue is the potential relevance of the correct development of trustworthy AI tools to reduce burn-out of healthcare professionals (correlating with adverse events) and medical malpractice (typically correlating with defensive medicine).



AI could really become an operational tool to manage these critical situations, by optimizing the role of human agents and their liability, thanks to decision-support systems and rigorous, standardised process data tracking. This issue could be very relevant in the short term for the healthcare domain, much more than more radically innovative solutions for diagnoses and therapies.

Additionally, education and training programs for health professionals must, in their various curricula, include a substantive discussion of AI, its promise and potential perils, and management of its risks.

Education for healthcare personnel, whether administration or physician/nurses, is a clear priority, too. Because the public struggles to understand the basics of how AI systems operate (Coeckelbergh, 2019), it should be assumed the same for healthcare personnel. In order for healthcare personnel to understand the risks inherent in use and implementation of AI systems in healthcare, they must be knowledgeable about these systems work - how algorithms arrive at specific decisions, how machine learning and big data can make predictive healthcare diagnoses. Educational literacy about AI for healthcare personnel could follow existing gamification models of digital educational training (Mesko et al., 2015).

Making sure that it is patients who benefit the most from the surge of AI health technology remains a key challenge. This will need new approaches in medical education to improve digital literacy, understanding of mathematical modelling, basis of decision theory, and continuous learning about AI of physicians. This should include awareness of biases in data, and how these undermine any claims about how AI models are able to produce objective, neutral results.

Accountability for AI systems in healthcare is also of great concern. The ability to audit healthcare systems is necessary, and could be built on the aforementioned models utilized in Helsinki and Amsterdam (Johnson, 2020), envisioning physician and patient ability to peer into the “black box” for auditability purposes. Further underscoring the need for education, audibility of AI systems is only possible when stakeholders involved in auditing these very systems comprehend the underlying technologies - where physicians do not fully comprehend all processes involved in AI (Diprose et al., 2020). Ethical-by-design healthcare AI needs to better integrate patients’ views and values to understand better different realities and kinds of knowledge, including the subjective aspect of illness. Patients’ wishes are a crucial measure for anticipating how AI technologies contribute to their health and wellbeing. Engineers and physicians need to work with patients to establish whether the use of AI is an empowered choice. This will need research programs to understand the patient’s own relationship with AI. A first step will be education and allow a better patient’s literacy. A second step will be



patient's engagement by feeding the dialogue with AI designers. The final step should be patient's empowerment to gain a better health through self-customized AI use.

Diversity for AI systems in healthcare must be also focused by industry and stakeholders from different perspectives, mainly such as:

1) Diversity of the team of the designers and developers of the AI-based solutions. As a minimum requirement the team should be balanced in gender, so to elucidate the wide spectrum of the needs, behavioural, communication and emotional styles which can be very different for male and female healthcare professionals, patients, relatives and all other human agents involved in the application scenarios

2) AI for diversity, i.e. the implementation of AI-based solutions which should cover the above general specifications by fully exploiting AI also for simulating the wide variety of diversity-open application scenarios, e.g. different facial morphologies and colors, different voice languages, expressiveness and accents, different motor behaviours, different cultural and social contexts. Using AI-based simulations can simplify, accelerate and better calibrate the development process so to deliver highly inclusive solutions.

3) Diversity of the sample population in data. The validation of the proposed AI-based solutions must be carried out by recruiting a diverse set of individuals so as to rigorously assess the actual performance when interacting with different\diverse human agents.

The current pandemic has increased examination of issues related to accessibility of healthcare. The implementation of AI systems in healthcare also brings forth issues of access, where we are now faced with scenarios of affordability. For example, if a private company markets an algorithm for detecting early onset of stroke, how can we ensure all governments have equal access and the technology is not out of reach economically? Rapid developments in AI will indeed increase issues of affordability and access. AI can become a key driver for the development of affordable healthcare solutions, optimizing cost-effectiveness, quality and dependability of novel solutions.

Privacy and security of data are important, as well. Machine learning requires massive amounts of data (Hedlund et al., 2020), and healthcare data possess a higher level of sensitivity and risk versus non-healthcare data. The security of these data and protection of patient privacy is imperative - however, we should not assume that GDPR is flexible enough to keep pace with seemingly quick developments in AI.

Perhaps our most important recommendation is that healthcare organizations need to design an explicit process for assessing AI risk and for mitigating that risk for each application of AI they are considering or using. That process must include the professionals, the organizational leadership, the patients and the public.



References

- Anderson T. The theory and practice of online learning. Edmonton, Canada: AU Press, 2008.
- Antun V, Renna F, Poon C, et al. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences* 2020: 201907377. DOI: 10.1073/pnas.1907377117.
- Arago, François. Speech before the French Parliament (Chambre des Députés), 14 June 1836, cited in *Le Monde*, 18 June 1954.
- Avizienis Algirdas , Laprie Jean-Claude, Randell Brian , and Carl Landwehr. Basic Concepts and Taxonomy of Dependable and Secure Computing. *IEEE Transactions On Dependable And Secure Computing*, Vol. 1, No. 1, January-march 2004.
- Barredo Arrieta Alejandro, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*, Volume 58, 2020, Pages 82-115, ISSN 1566-2535, <https://doi.org/10.1016/j.inffus.2019.12.012>.
- Bulla 2020. Bulla, Chetan & Parushetti, Chinmay & Teli, Akshata & Aski, Samiksha & Koppad, Sachin. (2020). A Review of AI Based Medical Assistant Chatbot. 2. 1-14. 10.5281/zenodo.3902215
- Char DS, Shah NH and Magnus D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *New England Journal of Medicine* 2018; 378: 981-983. DOI: 10.1056/NEJMp1714229.
- Castelvecchi, Davide. Can we open the black box of AI? *Nature* 538, 20–23 (06 October 2016) doi:10.1038/538020a
- Coeckelbergh M. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics* 2019. DOI: 10.1007/s11948-019-00146-8.
- Diprose William K. , Nicholas Buist, Ning Hua, Quentin Thurier, George Shand, Reece Robinson, Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator, *Journal of the American Medical Informatics Association*, Volume 27, Issue 4, April 2020, Pages 592–600, <https://doi.org/10.1093/jamia/ocz229>
- Eaneff, S, Obermeyer, Z and Butte, AJ The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *JAMA*. Published online September 14, 2020. DOI:10.1001/jama.2020.9371.
- European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry (COCIR). Artificial Intelligence in EU Medical Device Legislation. September 2020. https://www.cocir.org/fileadmin/Position_Papers_2020/COCIR_Analysis_on_AI_in_medical_Device_Legislation_-_Sept._2020_-_Final_2.pdf
- European Commission 2019a. Independent High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Brussels, 8.04.2019. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- European Commission 2019b. Independent High-Level Expert Group on Artificial Intelligence. Policy and Investments Recommendations. Brussels. 26.06.2019. <https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendationstrustworthy-artificial-intelligence>
- European Commission 2020a. White Paper On Artificial Intelligence - A European approach to excellence and trust. Brussels, 19.2.2020. COM(2020) 65 final.



https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligencefeb2020_en.pdf

European Commission. 2020b. Independent High-Level Expert Group on Artificial Intelligence. The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. Brussels, 17.9.2020. <https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificialintelligence-altai-self-assessment>

Featherstone K, Atkinson P, Bharadwaj A, et al. Risky Relations: Family, Kinship and the New Genetics. New York, NY: Taylor & Francis, 2020.

Floridi Luciano, Cows Josh, Beltrametti Monica, Chatila Raja, Chazerand Patrice, Dignum Virginia, Luetge Christoph, Madelin Robert, Pagallo Ugo, Rossi Francesca, Schafer Burkhard Valcke Peggy and Vayena Effy. AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, December 2018. Available at SSRN: <https://ssrn.com/abstract=3284141>

Grother Patrick, Mei Ngan, Kayee Hanaoka. Face Recognition Vendor Test (FRVT). Part 3: Demographic Effects. NISTIR 8280. Information Access Division Information Technology Laboratory, National Institute of Standards and Technology (NIST), 2019. Available at <https://doi.org/10.6028/NIST.IR.8280>

Grandjean Nathalie, Matthieu Cornélis and Claire Lobet-Maris. Sociological and Ethical Issues in Facial Recognition Systems: Exploring the Possibilities for Improved Critical Assessments of Technologies? 10th IEEE International Symposium on Multimedia, 2008.

Greene Gretchen. The Ethics of AI and Emotional Intelligence. Partnership on AI, 2019. <https://www.partnershiponai.org/the-ethics-of-ai-and-emotional-intelligence/>

Hedlund, J., Eklund, A. & Lundström, C. Key insights in the AIDA community policy on sharing of clinical imaging data for research in Sweden. Sci Data 7, 331 (2020). <https://doi.org/10.1038/s41597-020-00674-0>

Higher Health Authority, Paris, France, 2020 (in French). https://www.hassante.fr/upload/docs/application/pdf/2016-01/guide_fabricant_2016_01_11_cnedimts_vd.pdf.

Johnson K. Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI. Venture Beat, 2020.

Young-Bum Kim, Dongchan Kim, Anjishnu Kumar, Ruhi Sarikaya. Efficient Large-Scale Domain Classification with Personalized Attention. 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, July 2018.

King, D.. "DeepMind's health team joins Google Health" [Blog]. <https://deepmind.com/blog/announcements/deepmind-health-joins-google-health> (2019, September).

Lotman EM and Viigimaa M. Digital Health in Cardiology: The Estonian Perspective. Cardiology 2020; 145: 21-26. DOI: 10.1159/000504564.

Maurer D. Face Recognition Technology: DOJ and FBI Need to Take Additional Actions to Ensure Privacy and Accuracy. In: Justice HSa, (ed.). 2017.

McCall B. COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread, The Lancet Digital Health, vol. 2 e166-167, April 2020.

Merler M, Ratha N, Feris RS, et al. Diversity in faces. arXiv preprint arXiv:190110436 2019.

Mesko, B., Györfy, Z., & Kollár, J. (2015). Digital Literacy in the Medical Curriculum: A Course With Social Media Tools and Gamification. JMIR Medical Education, 1(2), e6. doi:10.2196/mededu.4411



Milani L, Leitsalu L and Metspalu A. An epidemiological perspective of personalized medicine: the Estonian experience. *J Intern Med* 2015; 277: 188-200. DOI: 10.1111/joim.12320.

Morley, J and Floridi, L. Policymakers must start asking difficult questions on the ethics of AI in healthcare. September 9, 2020. Available at: <https://www.publictechnology.net/articles/opinion/policymakers-must-start-asking-difficultquestions-ethics-ai-healthcare>. Accessed September 14, 2020.

Morley, J, Machado, CCV, Burr, C, et al. The ethics of AI in Healthcare: A mapping review. *Social Science & Medicine* 2020; 260: 113172. DOI:10.1016/j.socscimed.2020.113172.

Pagallo Ugo, Aurucci Paola, Casanovas Pompeu, Chatila, Raja, Chazerand Patrice, Dignum Virginia, Luetge Christoph, Madelin Robert, Schafer Burkhard and Valcke, Peggy.. AI4People - On Good AI Governance: 14 Priority Actions, a S.M.A.R.T. Model of Governance, and a Regulatory Toolbox (November 6, 2019). *A I 4 P E O P L E*, 2019, Available at SSRN: <https://ssrn.com/abstract=3486508>

Panch T, Mattie H and Celi LA. The “inconvenient truth” about AI in healthcare. *npj Digital Medicine* 2019; 2: 77. DOI: 10.1038/s41746-019-0155-4.

Philip Pierre, Lucile Dupuy, Marc Auriacombe, Fushia Serre, Etienne de Sevin, Alain Sauteraud, Jean-Arthur Micoulaud-Franchi. Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients. *npj Digit. Med.* 2020-01-07. 3(1)

Raji ID, Smart A, White RN, et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Fat** '20 2020: 33–44. DOI: 10.1145/3351095.3372873.

Shortliffe, E. H. (1999). The evolution of electronic medical records. *Academic Medicine*, 74(4), 414–419. Retrieved from <https://pdfs.semanticscholar.org/d46d/1c4f5871d3c915d220c7e0350c2c7054583b.pdf> [PDF]

Ting Daniel S. W., Yong Liu, Philippe Burlina, Xinxing Xu, Neil M. Bressler and Tien Y. Wong. AI for medical imaging goes deep. *Nature Medicine* · May 2018 DOI: 10.1038/s41591-018-0029-3

Vezyridis, P., & Timmons, S. (2017). Understanding the care.data conundrum: New information flows for economic growth. *Big Data & Society*, 4(1), 2053951716688490. doi:10.1177/2053951716688490

Vinuesa R, Azizpour H, Leite I, et al. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications* 2020; 11: 233. DOI: 10.1038/s41467-019-14108-y.

Weizenbaum, J. ELIZA: A Computer Program for the study of Natural Language Communication between Man and Machine, *CACM*, Vol. 9, Issue 1, January 1966

World Health Organization (WHO), Road Traffic Injuries (February 2020), <https://www.who.int/newsroom/fact-sheets/detail/road-trafficinjuries#:~:text=Approximately%201.35%20million%20people%20die,road%20traffic%20crashes%20by%202020>

Appendix: 7 Key Requirements for Trustworthy AI (European Commission 2019a).

Human agency and oversight

AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which



can be achieved through human-in-the-loop, human-on-the-loop, and human-incommand approaches

Technical Robustness and safety

AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.

Privacy and data governance

Besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimized access to data.

Transparency

The data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

Diversity, non-discrimination and fairness

Unfair bias must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.

Societal and environmental well-being

AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered.

Accountability

Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate and accessible redress should be ensured.

