ALLAI.

21.04.2022

RESPONSIBLE AI IN TIMES OF CORONA

This report is part of the Responsible AI & Corona Project.

Authors:
Catelijne Muller, LL.M
Mónica Férnandez Peñalver,
Christofer Talvitie, Laura Meyer,
Rosa van Ree

CONTENTS

1

INTRODUCTION

The Responsible AI & Corona Project was initiated for a number of reasons[1], but predominantly to assess whether AI could indeed play a valuable and responsible role during this crisis. Whether AI applications could effectively contribute to the understanding of the spread of the virus, to the search for a vaccine and to the treatment of COVID-19. Whether it could effectively guide policy measures and gain valuable insights that could help tackle the challenges of the crisis. And, most importantly, whether it could do this in a responsible manner. Because, despite the urgency of the crisis, it remains important that AI is robust, effective, transparent and explainable, and that fundamental rights, inclusivity and ethics are respected, to ensure that AI actually helps in tackling the Corona-crisis without causing harm to society along the way.

In times of crisis, we tend to turn more quickly to ‘invasive’ technology and to ‘turn a blind eye’ when it comes to fundamental rights, ethical values and even effectiveness. The motto is often: no harm, no foul. We see false contradictions such as: we have to choose between ethics or health, between fundamental rights or the opening up of the economy. And it is often thought that an invasive technology that is harmful, will be dismantled after the crisis. The reality is often very different. If AI does not help, it may very well harm. And once invasive technologies and practices have been introduced, it seems to be very hard to dismantle them again.

In efforts to tackle the Corona-crisis, many public and private parties indeed deployed AI-driven applications, both for medical as well as for societal uses. Examples of such applications include disease prediction with AI, AI to help vaccine discovery, face mask detection cameras and remote surveillance of workers with AI. Many of these applications are deployed through a fast-tracked decision cycle often driven by ‘techno-solutionism’, with less ‘eye’ for the implications of these applications for fundamental rights, ethical principles and societal values. This denotes the belief that for all complex social and in this case epistemological problems a technological solution can be developed. Even if basic elements such as effectiveness are often uncertain, there is a tendency to believe that in desperate times all possible avenues for solutions should be exhausted. This while it is well known that when powerful technologies such as AI are used unwisely, they can have serious unintended consequences.

Given the increasing number of solutions and the stated intention of many governments and public organisations to implement AI-driven systems to solve or alleviate the effects of this pandemic, the evaluation of such solutions was very important. Deploying organisations and end users, need to have the means to measure the effects of the solutions to be able to trust them. Such evaluation needs to go further than the technical characteristics of the systems, and include means to evaluate their societal, ethical and legal impact.

As part of the Responsible AI & Corona project, a first outline of evaluation criteria was developed to assess the efficacy and the legal, ethical and societal impact of AI-applications that are used during the Corona pandemic to guide their responsible use: a Framework for Responsible AI in times of Corona. The Framework aims to help organisations perform a ‘quickscan’ of the AI-application they want to develop, procure, deploy or use to tackle the challenges of this crisis. The Framework is a self-assessment tool to quickly identify the relevant elements of responsible AI and the level of adherence to these elements. It will help determine the level of impact of the AI-application, and provide options to balance different tensions and interests.

The Framework was tested with the Municipality of Amsterdam and on multiple use cases, the latter simultaneously being displayed in the “AI & Corona Observatory”, which also aims to highlight responsible, useful, reliable, and safe AI & Corona applications, and draw attention to unsafe, irresponsible, or undesirable AI.

As such this project has been providing insights in current 'AI versus Corona' research and development, by examining the state-of-the-art and maturity, the scope of use and the ethical, legal and societal impact of the applications. It has been evaluating tensions and possible trade-offs between conflicting values, interests and goals, taking into consideration the unique circumstances of this crisis. It has provided guidance on how to make optimal use of AI in this crisis within acceptable boundaries.

Despite the harshness and tragedy of this crisis, it also provided the circumstances and the momentum to step up efforts to find truly innovative and valuable AI applications, and to simultaneously tackle the ethical, legal and societal questions that have up to now been dealt with mostly through high-level principles, in a more practical manner.

---

1 Muller, Dignum, Schöppl (2020): Why do we need responsible AI in times of Corona?

2

A FRAMEWORK FOR RESPONSIBLE AI IN TIMES OF CORONA

As a first step in this project, we translated the translates the *Ethics Guidelines for Trustworthy AI*[2] of the EU into a *Framework for Responsible AI in Times of Corona* (the "Framework")[3]. The Ethics Guidelines for Trustworthy AI provide a set of requirements for trustworthy AI which include: 1. Human agency and oversight; 2. Technical Robustness and Safety; 3. Privacy and Data Governance; 4. Transparency; 5. Diversity, Non-Discrimination, and Fairness; 6. Societal and Environmental Well-being; and 7. Accountability.

Using these requirements, the Framework was adapted to the circumstances of this crisis by looking at 3 relevant areas: **1. Impact on Citizens and Society; 2. Technological Robustness and Efficacy; 3. Governance and Accountability:**

1. Impact on citizens and society can be positive, where the use of the AI system for example leads to the improvement of health and safety, better access to information, better social well-being, or negative, where it leads to an undesired precedent for the future, adverse impact on human rights, bias, etc.
2. Technological robustness and efficacy looks at security, accuracy and reproducibility of the AI application, while efficacy looks for example at whether there is a sound problem definition and if the AI application contributes to solving the problem.
3. Governance looks at whether there is a legal basis (necessary) for the deployment of the AI application, and if for example a 'sunset clause' exists. It also looks at purpose limitation, i.e. a guarantee that the AI-application is only used to serve the specific purpose. Accountability looks at if there is a clear structure of responsibility.

Together these areas contain 24 requirements that should be taken into consideration when developing, deploying or using AI in a responsible manner during a crisis.[4]

SOCIETY	TECHNOLOGY	GOVERNANCE
1. Impact on Society	9. Problem Definition	17. Legal Basis
2. Human Rights	10. Solution Optimization	18. Necessity & Proportionality
3. Democracy	11. Effectiveness	19. Domain & Purpose Limitation
4. Legal Compliance	12. Adverse Effects	20. Transparency
5. Human Agency	13. Security	21. Voluntariness
6. Explicability	14. Accuracy	22. Sunset Clause
7. Fairness	15. Generalisability	23. Stakeholder Involvement
8. Accessibility	16. Human Oversight	24. Accountability

The Framework was tested through two 'routes'. First an evaluation of the Framework and the 24 requirements was performed with the AI-lead of the Chief Technology Office of the Municipality of Amsterdam. Second, the Framework was used to assess a total of 14 specific use cases, where AI was developed or deployed to tackle a challenge of the crisis. These use cases and assessments are being displayed in the AI & Corona Observatory. The testing phase has led to some minor adaptations of the Framework, which we describe in the final chapter of this paper.

2. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
3. https://allai.nl/wp-content/uploads/2021/01/A-Framework-for-responsible-AI-in-times-of-Corona-1.pdf
4. Ibid.

3

THE AI & CORONA OBSERVATORY

Through extensive desk research, a total of 14 AI applications were evaluated against the Framework. Apart from the 3 relevant areas (society, technology and governance), an indication of the scope of use of each AI application was given as well as an indication as to whether acceptable trade-offs between conflicting requirements could be found.

Under normal circumstances, tensions can rise between requirements that necessitate a careful balancing of interests and sometimes lead to the trade-off of one requirement against the other. The current extraordinary circumstances can lead to different tensions, new or different interests, a different balancing of those interests and thus different trade-offs when evaluating an AI application.

Under the extreme circumstances of this crisis however, the right governance measures (e.g. voluntariness, a sunset clause, a controlled environment) might tip the balance in favour of exceptional use of the application. In contrast, an application with a low level of efficacy could nevertheless be interesting to deploy for research purposes, if it poses no ethical, legal and societal risks.

We broadly distinguished two areas of AI use: 'medical' and 'societal'. Medical use of AI involves applications that are directly related to the medical aspects of the Corona-crisis: epidemiology, molecular research, clinical practice and provision of care. Societal use of AI involves applications that are related to the public and private aspects of the Corona crisis: policy measures (social distancing, mask wearing, lock-downs, etc.), distance learning, remote working, delivery of information, etc.

For most of the AI applications, a use case example allowed us to evaluate the application in more detail against all 24 requirements. In these cases, each requirement was given a score from ‘0’ to ‘3’ to demonstrate the following:

0 = unknown (no information was found to reach a conclusion about whether the application complies with the requirement)
1 = no compliance (the application does not comply with the requirement)
2 = partial compliance (the application complies partially with the requirement)
3 = full compliance (the application fully complies with the requirement)

The project highlights a total of 16 AI(-driven) applications of which we assessed a total of 13 against the Framework in the following areas: Checking and enforcing of Corona measures; COVID Risk detection and prediction; Education; Work; Healthcare; Information:

1. Face Mask Detection
2. Counting cameras
3. Remote Thermal Scanning
4. Movement Tracking
5. Population Vulnerability
6. Online Proctoring
7. Algorithmic Grading
8. AI & Home Office
9. Vaccine Discovery
10. Cough Detection
11. COVID diagnosis
12. COVID prognosis
13. Vaccine Hesitancy Chatbot

3.1 Face Mask Detection

ThroughThe onset of the pandemic has given rise to technology that monitors compliance with COVID-19 measures, such as AI-driven face mask detection. Many systems were being developed for this purpose, with a growing community of developers sharing their coding methods through GitHub (a code hosting platform). The scope of use of these systems was difficult to determine, but our research indicated that they are being tested and/or used often covertly.

Face mask detection was briefly tested at the **Paris Châtelet-Les Halles Metro station** by the French RATP before being cut short due to criticism, amongst others by the French privacy watchdog CNIL.

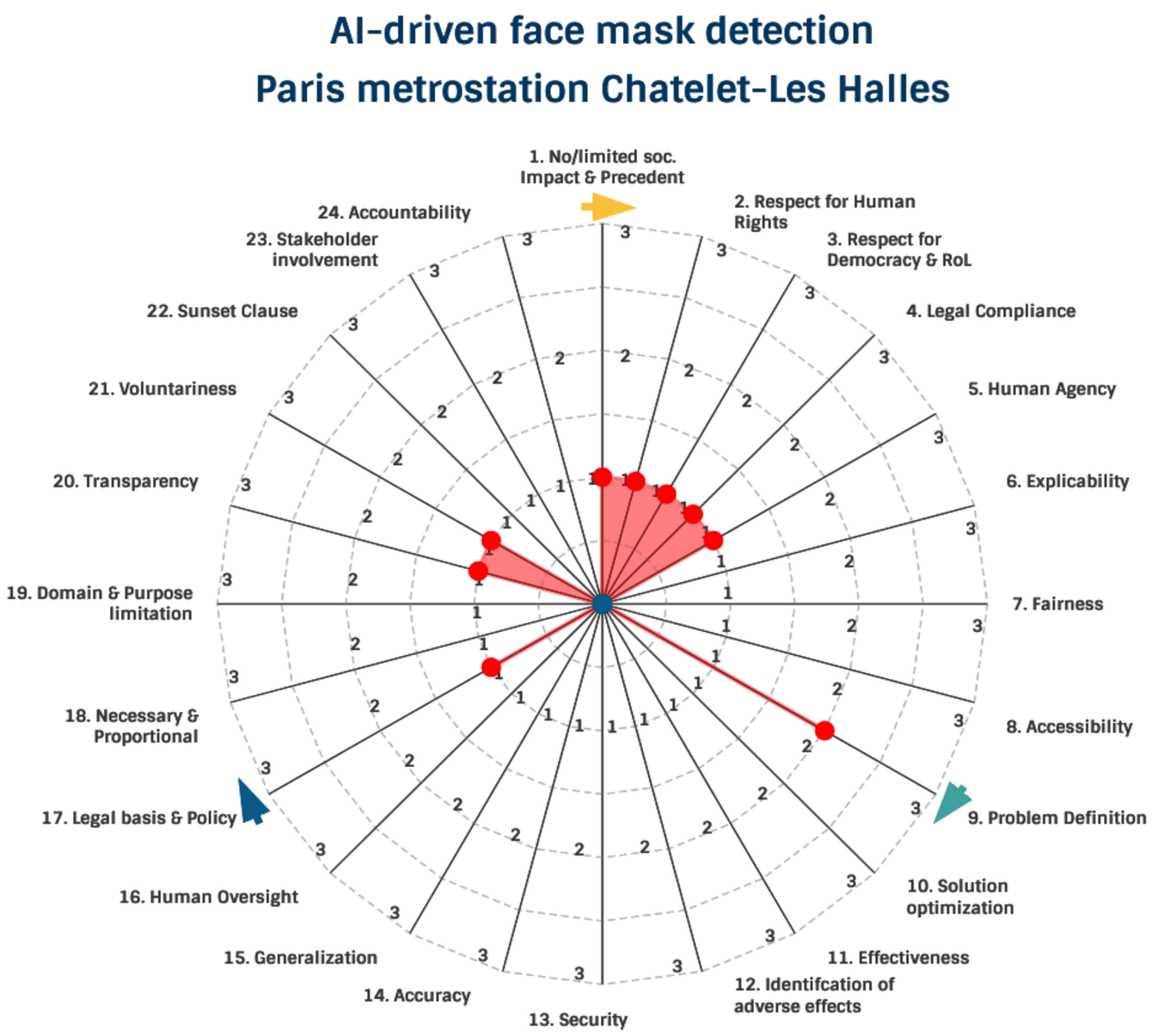
Like other forms of intrusive surveillance, face mask detectors have a **negative impact on citizens and society** because they set an undesired precedent for the future, because of normalisation and increased acceptance of surveillance. Furthermore, we found that these systems negatively impact the right to privacy, autonomy, human agency, and free will. It can create a 'chilling effect', where people adjust their behaviour to a certain norm because they are constantly being watched. CNIL noted the difficulty of complying with the GDPR, since people cannot give consent to their data being processed by the system.

We have concerns about the **technological robustness and efficacy** of the system due to a lack of generalizability. This means that it works well in a controlled (testing) environment (e.g. where a small number of white men, that are well lit, wear the same mask, and look straight into the camera, passing at the same pace), however, it is highly uncertain how it will perform when used in a real-life, crowded, poorly lit or fast-paced settings.

Our research found little information regarding **governance and accountability**, as the use and/or testing of these systems are currently 'under the radar'. Given the lack of publicly available documentation, it is unknown whether there is an appropriate legal basis that regulates the use of the system in times of Corona *i.e.,* the existence of a sunset clause.

The graph below displays the scores obtained for the Châtelet-Les Halles' case of AI-driven face mask detection. The results clearly show a lack of publicly available information to determine their compliance with most of the requirements. The only available information showed it only partially complies with *'problem definition' and* does not comply with 8 out of 24 requirements.

Given the intrusiveness of facial recognition applications and the fact that multiple governments and legislative bodies are developing or are calling for strict regulation or even a ban on facial recognition, we see **no acceptable trade off** at this stage for the use of face mask detection cameras.



3.2 People Counting Cameras

The Falling under the scope of surveillance technology, we have stumbled upon AI-driven counting cameras; cameras that aim to monitor the number of people in establishments to help enforce social distancing in public places. MOBOTIX, Hikvision, V-count and Canon are a few of the many companies offering these products. Although the deployment of counting cameras is being justified to fight the crisis, many of these technologies contain intrusive AI-based features that can negatively impact citizens and society.

We have chosen **Leiden university's deployment of counting cameras** as a use case example to analyse the impact of the technology. Leiden University has deployed 350 AI-driven 'counting sensors' manufactured by Xovis during the lockdown of early 2021. These counting sensors have been deployed in classrooms and corridors to monitor the number of people on university premises. However, the cameras can do much more than just count. They provide AI extensions such as gender statistics, gaze direction, face mask detection, staff exclusion, and group counting.[5] The amount of data that the user sees and how anonymous it is depends on these settings. Leiden university claims to be using it at a level where people are identifiable as silhouettes.[6] They guarantee that they do not register any specific characteristics of people. However, the university only informed their students and staff about their use after they had been exposed by Mare, the university's independent weekly magazine.[7]

As with any AI-based surveillance technology, counting cameras have shown to be unnecessarily intrusive. Their additional AI-based features exacerbate their negative **impact on citizens and society** by disregarding people's right to privacy and autonomy and creating an unavoidable 'chilling' effect on students and staff. Moreover, students and staff were not aware of the use until one year after their deployment. This left many of them feeling disrespected, intimidated, and fearful towards their non-consensual use.[8] The application only obtained full scores regarding fairness (as they do not show any sign of unfair bias) and accessibility (as they are used for the monitoring of all).

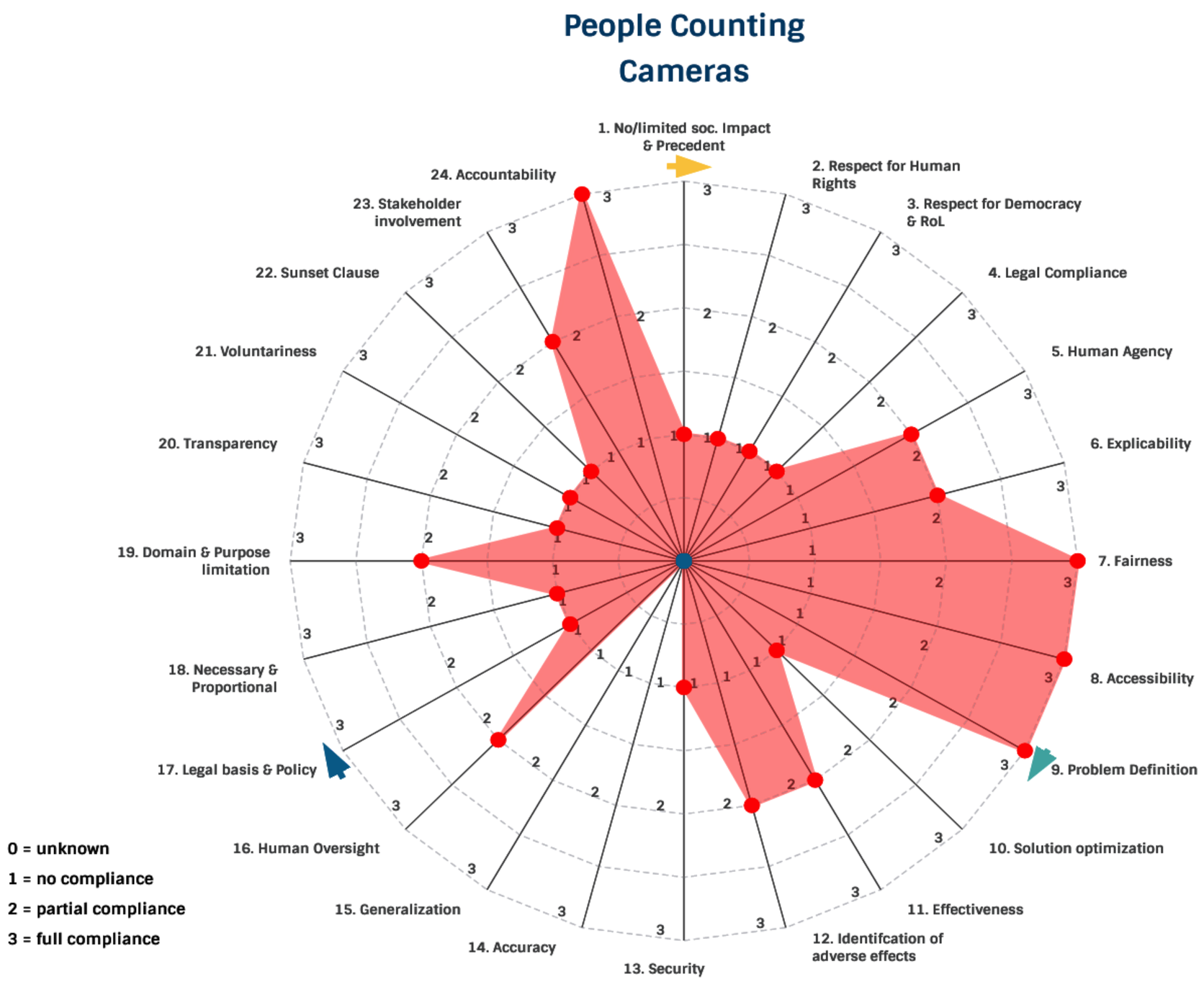
Our analysis showed mixed scores regarding its **technological efficiency and robustness**. Xovis claims that their cameras have a 99% accuracy at counting the number of people.[9] However, we have not found any evidence supporting this claim. Although Leiden university has claimed that these cameras were more efficient than translating WiFi signals or CO_2 levels, we have not seen any attempts at testing less intrusive, 'old-school' counting methods. Furthermore, whether measures are being taken in response to the outcome of the system is unknown. The system would only be efficient if appropriate measures are taken in response to the outcomes of the counting sensors. We have not found any procedural forms from Leiden University confirming this. Lastly, the technology did not score well in 'security', because cameras systems are an ideal target for hacking, and until recently, the login page of the cameras was unprotected via the public internet, and the data that the cameras collected was "protected" with an unencrypted password only.[10]

Regarding its **governance and accountability**, the university claims the existence of a processing agreement that data may not be used for any other purpose, except on the instruction of the university. This gives reasons to be concerned given the capabilities of the technology and the fact there was no clear, open and direct communication on their workings and use to students and staff for over a year. Submission to the cameras was obligatory, as the only way to avoid them was to not visit the university premises. Lastly, the deployment of these systems in public spaces without clear communication and transparency about the system's use and workings do not fall under the scope of some legal and policy frameworks such as the GDPR.

The figure on the next page displays the scores obtained through the case of People Counting Cameras deployed at Leiden University. The results clearly show a lack of compliance with 11 requirements, partial complience with 6 requirements, and full complience with only 3 requirements. Given their negative impact on citizens and their mental well-being, as well as the availability of less invasive old-school methods to count the number of people in a room, we see not acceptable trade off at this stage for the use of counting cameras in times of corona.

---

5. https://www.xovis.com/technology/sensor/pc2s-sensor
6. https://www.medewerkers.universiteitleiden.nl/mededelingen/2021/11/universiteit-telt-aanwezigheid-van-studenten-en-medewerkers-met-scanners-aan-plafond?cd=centre-for-linguistics#1-voldoen-de-classroom-scanners-aan-de-avg,2-nemen-de-classroom-scanners-mij-op,3-kunnen-de-classroom-scanners-mij-identificeren,4-welke-informatie-verzamelen-de-classroom-scanners
7. https://www.mareonline.nl/achtergrond/opeens-hangen-er-overal-slimme-cameras-en-die-zien-alles/
8. https://www.universiteitleiden.nl/en/news/2021/12/protest-against-scanners
9. https://www.xovis.com/technology/sensor
10. https://www.mareonline.nl/achtergrond/opeens-hangen-er-overal-slimme-cameras-en-die-zien-alles/



3.3 Thermal COVID-19 Risk Detectors

In order to prevent and contain the spread of COVID-19, thermal cameras have been widely used across airports, public transit hubs, health facilities and other public spaces, to detect elevated body temperature of individuals in crowds.[11] The collection of sensitive biometric data gave us reason to investigate these systems critically and make sure that it is responsibly used.

We took the case of **thermal cameras deployed across airports in Spain** as a use case to perform our assessment. Our research findings, however, show that these systems have a low level of **technological robustness and efficacy**, as they are susceptible to errors in temperature readings, particularly when they are used to scan multiple people in crowds. It is important to note that these cameras can only predict surface body temperature, not internal body temperature, thus only contributing to the detection of COVID-19 through indirect measures, and missing the detection of asymptomatic people, or those going through an incubation phase.[12]

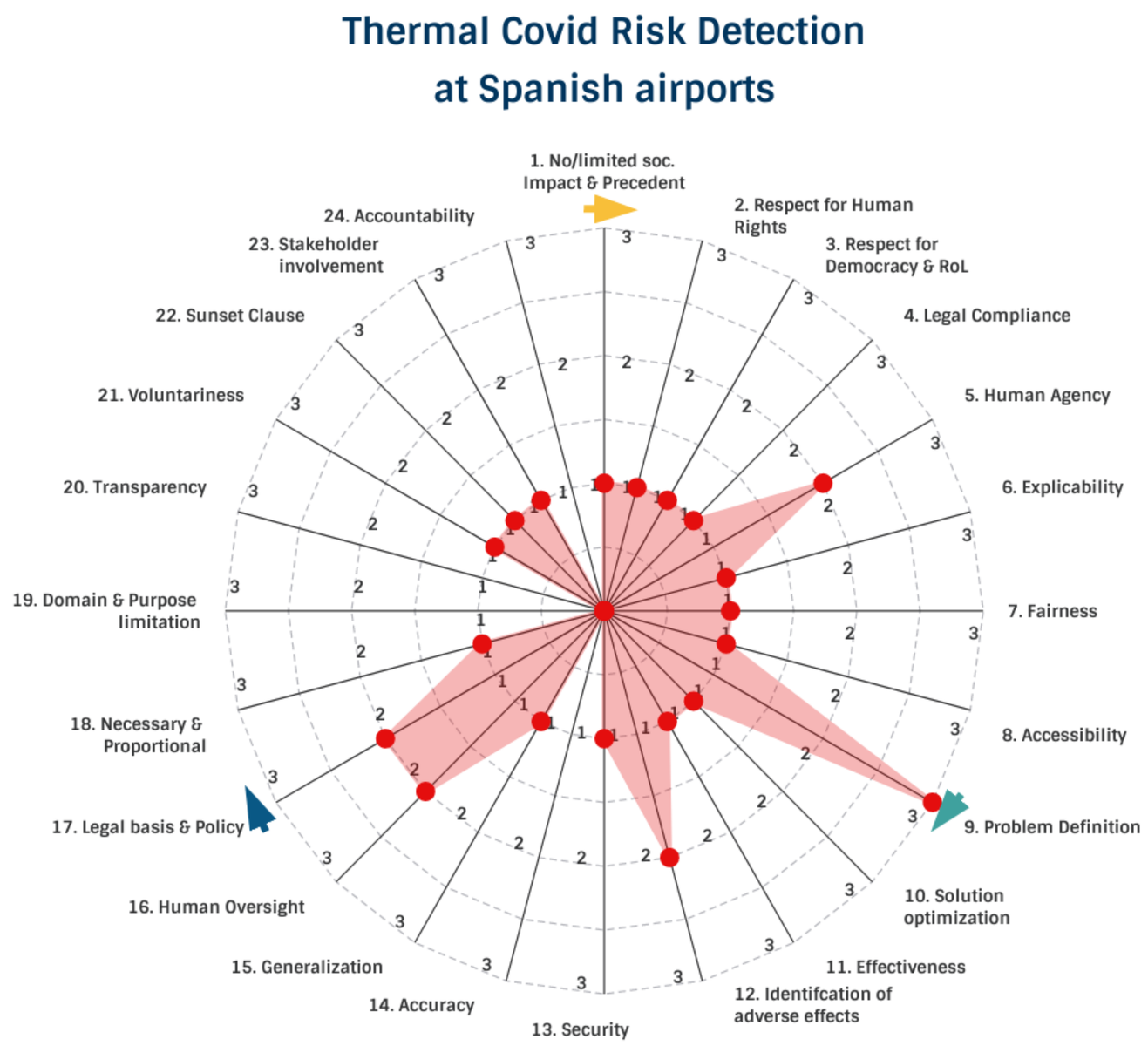
Even if the cameras were highly accurate and efficient, we have found that their use has a negative impact on **citizens and society** due to various concerns. In any context, they are a highly intrusive form of surveillance, that creates a chilling effect of constantly being surveilled. It also sets an undesired precedent for the future by normalising the collection of sensitive personal data. Furthermore, there is an ample impact on the human right to privacy, autonomy, and democracy.

Regarding its **governance and accountability**, the use of thermal cameras in public spaces is not in compliance with the GDPR, because people cannot give consent to the collection and processing of their personal biometric data. Although surveillance companies argue that the responsibility to comply with policy and legislation fall upon the controller of the AI system, there seems to be no clear and specific legislation put in place for each of these AI-based surveillance devices. Furthermore, it is unknown whether people's data is used for other purposes *i.e.*, training the AI or statistical analyses or whether additional data is collected aside from what is strictly necessary.

11. https://fpf.org/blog/thermal-imaging-as-pandemic-exit-strategy-limitations-use-cases-and-privacy-implications/
12. https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.5.2000080#html_fulltext

The use of thermal cameras has been justified by many European countries because of the seriousness of the pandemic. Yet, scientists argue that they are not better at detecting internal body temperature than alternatively, less-invasive solutions e.g. infrared thermometers.[13,14] Given the high cost of indirectly detecting people infected with covid-19 (whilst negatively impacting privacy, autonomy, and democracy) we see no **acceptable trade-offs** at this stage for the use of AI-based thermal cameras.

The figure below displays the scores obtained for Thermal Covid-Risk Detection deployed at Spanish airports. The results show that these applications only comply fully with '*problem definition*' and '*accessibility*'. They partially comply with 2 requirements: *human oversight,* and *stakeholder involvement* requirements. Nevertheless, the results clearly show that they do not comply with the rest of the requirements based on publicly available information.



3.4 Movement and Contact Tracking and Tracing

Since the start of the pandemic, 120 COVID contact-tracing apps have become available in 71 countries, and 60 digital tracking measures have been introduced in 38 countries. Contract-tracing apps and digital tracking measures aim to track people's locations or who they have been in contact with through GPS or mobile Bluetooth methods. Although they all share a similar purpose, these applications differ widely in their intrusiveness, what sort of data they collect, whether they are (de-) centralised, which technology they are based on, whether they are mandatory and whether they are managed by the state.

We have chosen the case of **South Korea's centralised movement tracking system**. South Korea has adopted the use of COVID contact tracing apps to track and monitor the spread of the virus among its citizens. Several private and public organisations assess the data such as GPS data, times and locations and travel routes. The data is given to the Korea Centres for Disease Control and Prevention (KCDC) to be published (without personal identifiers) on the government's website.

13. Ibid
14. https://www.who.int/news-room/articles-detail/key-considerations-for-repatriation-and-quarantine-of-travellers-in-relation-to-the-outbreak-of-novel-coronavirus-2019-ncov

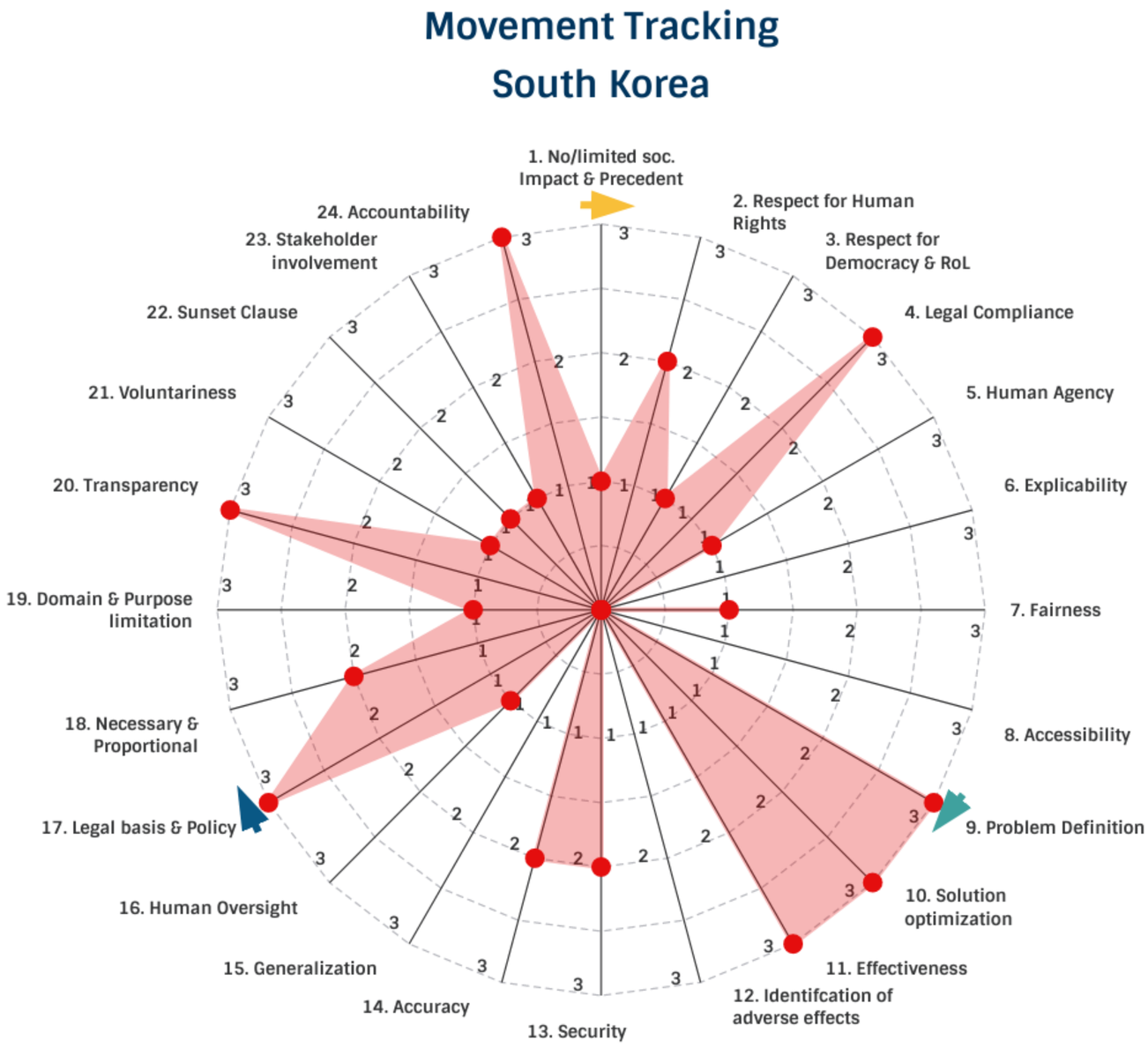
SOCIETAL

Regarding its **technological robustness and efficacy**, research shows that centralised movement tracking apps which monitor the movement of the citizens are the most efficient anti-contagion method. This method has allowed Korea to control and detect the early spread of the virus way better than other countries of similar population sizes. In some cases, however, the process unveiled or inferred embarrassing personal details of individuals.

Regarding its **impact on citizens and society**, the common criticism is that movement tracking and tracing will lead to a permanent change in societal habits and behaviours and an altered level of trust in government. Our research has found that submission to being tracked is usually not voluntary, indicating the violation of the individual's right to privacy. And as with other forms of surveillance, movement tracking sets a dangerous precedent for ubiquitous surveillance, infringing both individual and collective rights and the fabric of society.

This case clearly shows the dilemma with all contingency measures: the **trade-off** between intrusiveness and efficiency. Despite its intrusiveness, its high effectiveness has led to many national legislations being altered during the Corona crisis to allow movement tracking and monitor quarantines. Yet, we have not been able to determine whether appropriate democratic processes were followed and whether the different legislations provide for a sunset clause.

The figure below displays the scores obtained for the use of Movement Tracking systems in South Korea. The results show that these applications comply fully with 6 requirements: '*problem definition*', '*solution optimization*', '*effectiveness*', '*legal basis & policy*', '*transparency*', '*accountability*', and *0legal compliance*'. They partially comply with 4 requirements: '*respect for human rights*', '*security*', '*accuracy*', *and* '*necessity and proportionality*', and do not comply with 9 out of 24 requirements (the rest are unknown).



0 = unknown

1 = no compliance

2 = partial compliance

3 = full compliance

3.5 Algorithmic Grading

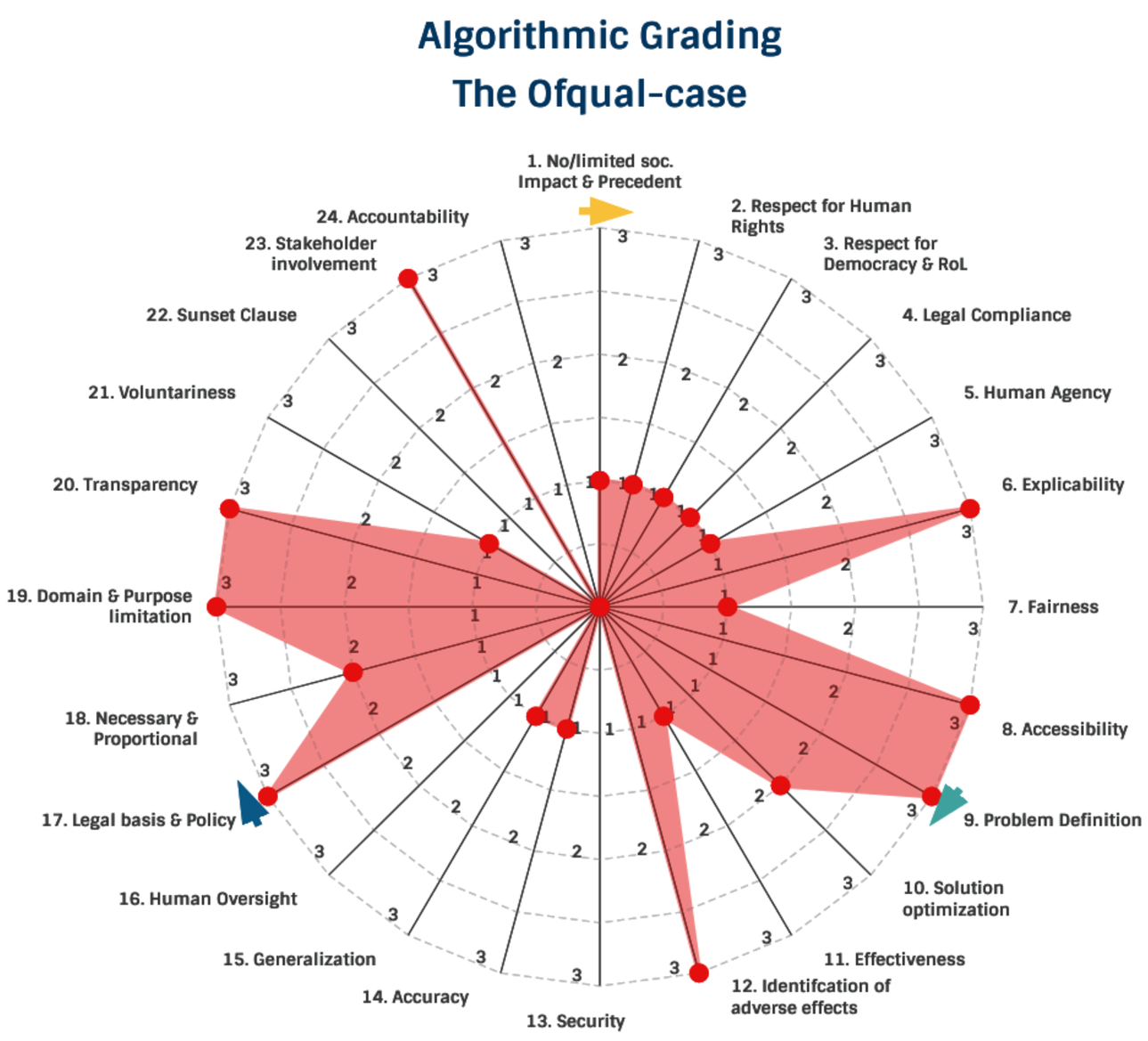
The onset of the pandemic in 2020 has led many educational institutions to adopt the use of algorithmic grading to replace end-of-year examinations such as the International Baccalaureate (IB) final examinations or the A-level examinations in the United Kingdom. It is estimated that over 1,000,000 students worldwide have had their future education/career be determined by a predicted grade. These predicted grades were based on students' past work, teacher's predicted grade, the past success of the school and ranking of the students within the school.

As a use case we chose the Ofqual grading algorithm in the UK. The use of algorithmic grading has been highly criticised due to the **low technological robustness and efficacy** that these systems provide. The algorithmic models usually are set to predict the likely distribution of grades, not assessing the performance of a student on his/her own merits, making it unjust. By looking closer at the Ofqual (Office of Qualifications and Examinations Regulation) Grading Algorithm used in the UK, our research has shown that in the test phase, where the results of the algorithm were compared with actual grades, the accuracy of the model used for the "A-levels algorithm" was low – in the range of 50 – 60%.

Given the low accuracy and design choices of the models, the applications showed to have a **negative impact on citizens and society**. Unjust or incorrect grading detrimentally affects the life opportunities of a student. Regarding its design, such as the inputs fed into the system, using school successes as variables unfairly favours better (and usually private) schools over less successful (usually public) schools, exacerbating inequality.

Research on **governance and accountability** showed that a vast governance structure was set up, involving key stakeholders such as teachers, students, and student's parents. However, there was no possible voluntary submission to the algorithmic grading and little information on accountability was found.

The figure below displays the scores obtained for the Ofqual case of algorithmic grading. The results show full compliance with some requirements concerning governance; *'legal basis & policy'*, *'domain & purpose limitation'*, *'transparency'*, and *'stakeholder involvement'*. Outside this area, the application complied with the *'identification of adverse effects'*, *'explicability'*, *'accessibility'*, and *'problem definition'*. Furthermore, it partially complies with the '*necessity and proportionality'* requirement, and '*solution optimization'* requirement. The rest of the requirements were not complied with (or no information was found regarding its compliance).



3.6 AI-driven Proctoring

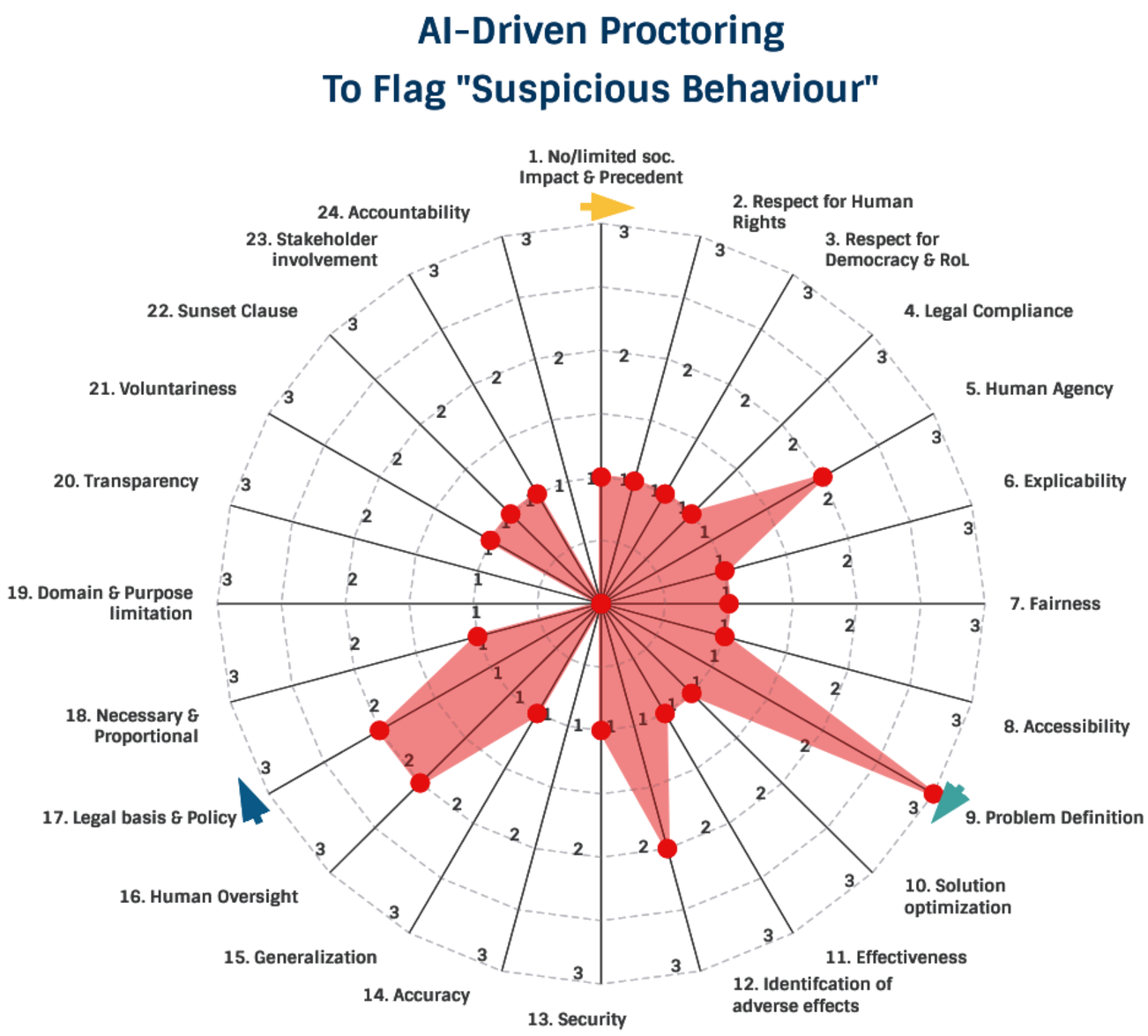
Because of extensive lock downs, many educational institutions decided to have students take their exams remotely, in order to guarantee the continuation of education. In order to perform surveillance, many universities adopted the use of AI-based online proctoring, which supposedly recognises 'suspicious behaviour' and reports instances of potential fraud during online examinations. The aim of online proctoring is to preserve academic integrity and honesty in online exams. However, use of online proctoring has faced objections from student-representative bodies who deem online proctoring to infringe privacy and autonomy of students by exposing them to extensive surveillance. Many online proctoring systems detect fraud by what they believe to be “excessive looking away from the screen”, open webpages, on-screen activity and open personal files during exams. Furthermore, students are required to “360 scan” their room and keep their web-camera on during the exam, exposing the private space of the students at home. AI analysis claims to identify suspicious examinee behaviours or suspicious items in their immediate physical or digital environment.

As regards **scope of use**, a 2020 poll found that 54% educational institutions used online proctoring and 23% are considering its use. The actual scope of use was hard to define. However, with respect to its **technological robustness and efficiency**, no extensive research is available confirming that AI-based proctoring is more effective than non-AI-based proctoring at detecting fraud, suggesting it might not be the optimal solution to tackle the problem. Furthermore, there have been cases of inaccurate detection of fraud, and it is not clear whether the system is fully generalizable to different situations outside its training environment.

Regarding its **impact on citizens and society**, online proctoring shows to be a highly intrusive form of surveillance that does not respect the student’s right to privacy, mental integrity, or autonomy. While it is argued that online proctoring ensures academic integrity and honesty in online examinations, it might set an undesired precedent for the future because of normalisation and a higher acceptance of intrusive surveillance. Furthermore, it deteriorates the culture of mutual trust between students and teachers that prevails in educational institutions.

Regarding **governance and accountability**, AI-based proctoring shows elements of non-compliance with existing laws such as the GDPR, as consent cannot be considered freely given by the students. Although the methodology (i.e., a 360° scan of the room, access to web pages visited and personal files, and constant monitoring of eye and/or head movement) is already deemed as unnecessary and disproportionate to achieve the desired result, it is also unknown whether student’s data is used for other purposes such as training the system. Lastly, in several cases, student bodies were not involved in the decision to apply the system, showing an undemocratic practice regarding its implementation.

Online proctoring is justified based on preserving academic integrity and honesty amid the pandemic. However, it is not clear to which extent AI-based online proctoring is better at detecting fraud than non-AI-based online proctoring. AI-based proctoring has a significant **trade-off** on student’s privacy, autonomy, and places student’s personal data into a vulnerable position. It is also not clear to which extent access to student’s personal files and web-page visits or monitoring student’s eye-movement is necessary for detecting fraud in examinations.



3.7 AI & the Home Office

Prior to the pandemic, 10% of businesses had adopted applications for home office monitoring to ensure employee productivity, but this number increased to 30% over the past two years. Home office monitoring systems include applications such as Office 365 Monitoring, ActiveTrak, and Teaming, among others. Many of which invasively use eye-tracking, location tracking, mouse movement, comply and paste behaviour, search behaviour, soundtracking etc.

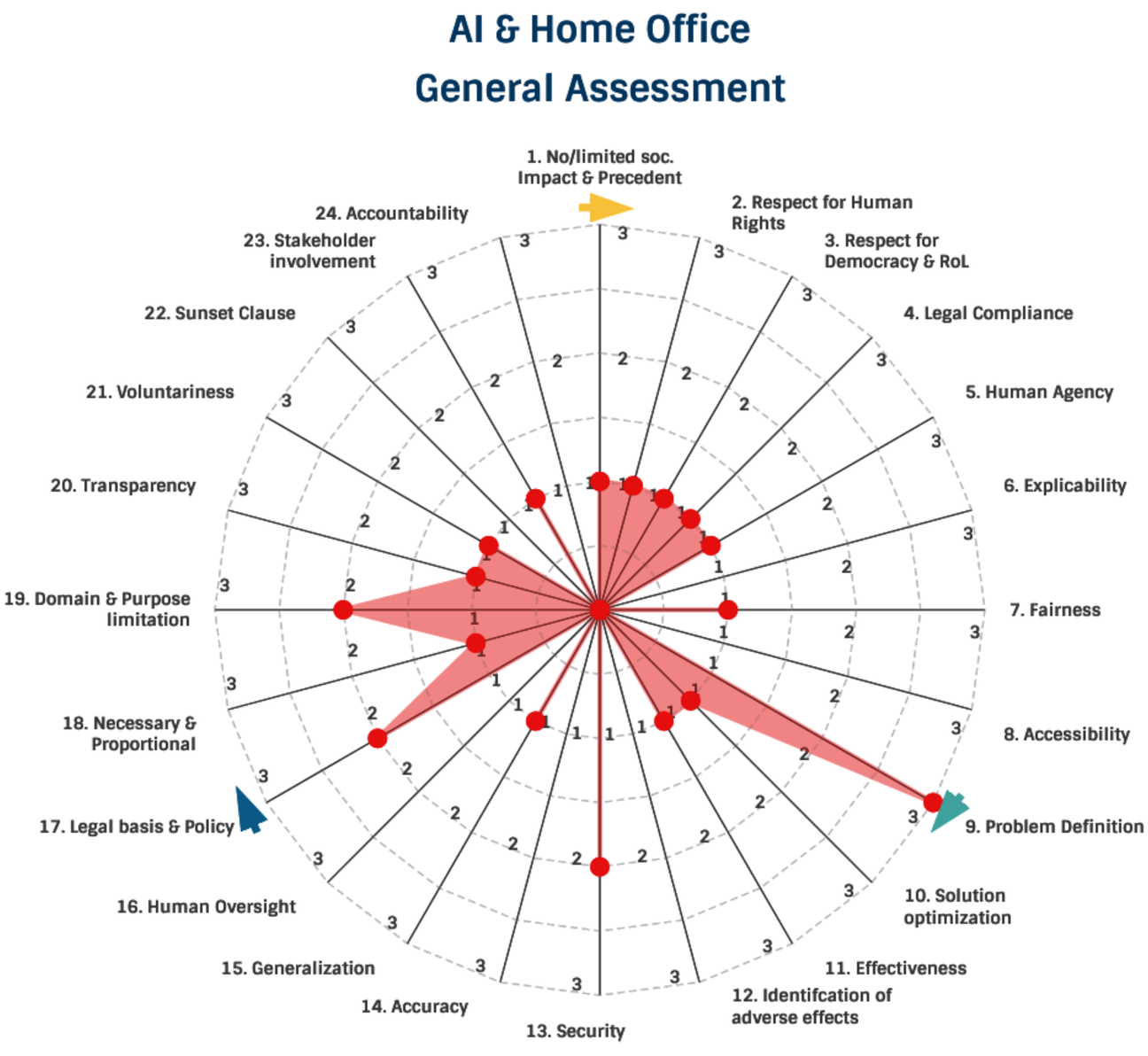
Despite the invasive tracking elements that these systems use, they have shown a **low level of technological robustness and efficacy**. For example, there is no research that shows that these AI-based monitoring tools are more effective than non-AI-based approaches. It is also unclear whether the features being monitored are representative of true productivity and performance (these features also become even more difficult to define with more complex/creative positions). Moreover, it has been shown that the systems can be outsmarted in different ways raising concerns about accuracy and fairness.

Similarly to online proctoring, home office monitoring has a **negative impact on citizens and society**, as constant surveillance can undermine the trust, confidence and well-being of employees. It affects the right privacy, undermines human agency, and might set the undesired precedent: more acceptance of intrusive surveillance.

Regarding **governance and accountability**, our research suggests that depending on the country, there might be national labour laws that require a specific legal basis for worker monitoring, and collective bargaining agreements could have specific rules. Yet, most of the time employees are not involved in the decision of applying the system and their submission to it is not voluntary (either because employees are not made aware of the use of the system, or they are obliged to consent).

We see no acceptable **trade-offs** in times of crisis, primarily because it is not clear whether monitoring tools have a positive or a negative impact on employee's productivity and performance. Policies to support the transition to more widespread remote work will need to carefully consider the potential benefits and costs for productivity, job quality, and workers' work-life balance and mental health. This is especially important in view of rising demand and development of home office monitoring systems that will likely not be scaled back after the end of the pandemic.

Figure 8 displays the scores obtained for general AI-based home office monitoring systems. The results show that these applications only comply fully with '*problem definition*'. They partially comply with 3 requirements: *security*, *legal basis & policy*, and *domain & purpose limitation* requirements. Yet, it is highly noticeable that they do not comply with most of the requirements based on publicly available information.



MEDICAL

3.8 AI & Vaccine Discovery

AI models to assist in vaccine discovery and development remain largely untested and are mainly theoretical. Nevertheless, one of the most direct applications of AI is to identify the presence of antigenic peptides presented by MHC-II (molecules that induce antigen-specific responses, which are central to vaccine-induced immunity). The study of this peptide in patients is aimed to understand natural immunity and discover COVID-19 specific immune responses that can assist in designing effective vaccines. Other AI tools researched to predict antigen presentation include MARIA, NetMHCPan4, Long Short-Term Memory network, deep-learning Recurrent Neural Networks and MoDec among others.[15]

In recent years the successful application of machine learning has revolutionised many fields of science including vaccine discovery. However, in the context of COVID-19 vaccine discovery and development, there is no adequate evidence regarding their **robustness and efficacy**.

We do consider this type of AI to have a predominantly **positive impact on citizens and society**, given its potential to save lives against deadly viruses. However, as in any medical and public health setting subjected to AI-based decisions, validation, generalisation, explainability, interpretability, risk mitigation, fairness, and inclusiveness are some of the key challenges that should be addressed to protect citizens and society from any adverse effects.

We believe that adequate **governance and accountability** is of special importance in clinical and healthcare settings, especially during a pandemic. Transparency, privacy, fairness, safety, and liability should be prioritised. This can be done in different ways. For example, in the context of COVID-19 vaccine discovery, issues concerning bias and lack of transparency should be dealt with by engaging different stakeholders during the process. Furthermore, explainability and interpretability are two important factors that need governance to monitor and enhance AI algorithmic fairness, transparency, and accountability. Lastly, “ethical auditing” could be deployed to examine the inputs and outputs of AI algorithms and models for bias and potential risks that AI could bring upon any application.

3.9 AI & Population Vulnerability Prediction

Many studies have revealed that health inequalities amongst populations such as front-line workers, marginalised communities, and racial minority groups face a higher risk of morbidity against COVID-19.[16, 17, 18, 19, 20] Part of this vulnerability is linked to the higher risk of having comorbidities (e.g. high BMI, respiratory diseases) that have also been defined as major risk factors of illness and death from COVID-19. Researchers have therefore been working on developing AI models that can identify people in the general population at risk of covid-19 infection or at risk of being admitted to hospital with the disease These models may work by taking multiple comorbidities or social factors (e.g. demographics occupation *etc.*) into account.

The methodologies differ widely. One researched model uses thermal cameras to detect abnormal breathing through face recognition techniques. Other models use medical datasets *(i.e.* hospital admission for non-tuberculosis pneumonia, influenza, acute bronchitis *etc.)* as proxy outcomes to determine vulnerability . Lastly, another model used demographics, symptoms, and contact history in a mobile app to assist general practitioners in collecting data and to risk-stratify patients.

Given the complexity of the interplay between social factors and comorbidities, predicting population vulnerability is not an easy task to do. Although the models mentioned above claim to have high accuracy index, the PRECISE living review revealed that all the models mentioned above had an unclear or high risk of bias regarding the use of participants, outcome, and analysis.

Given that most of the models found were not developed wisely, had poor methodology or displayed technical biases, we consider this type of AI to have a predominantly negative **impact on citizens and society** at this point. Using models that have a high risk of bias can exacerbate health inequality and other social inequalities. In times of corona, this can even lead to harmful or fatal consequences. Furthermore, the use of invasive techniques such as facial recognition and biometric identification in public places entail further negative impacts on citizens and society. (see assessments on covid risk detection or face mask detection to learn more).

---

15. Keshavarzi Arshadi, A., Webb, J., Salem, M., Cruz, E., Calad-Thomson, S., Ghadirian, N., ... & Yuan, J. S. (2020). Artificial intelligence for COVID-19 drug discovery and vaccine development. *Frontiers in Artificial Intelligence*, *3*, 65.
16. Lainjo, B. (2021). The Enigmatic COVID-19 Vulnerabilities and the Invaluable Artificial Intelligence (AI). *Journal of multidisciplinary healthcare*, *14*, 2361.
17. Patel, J. A., Nielsen, F. B. H., Badiani, A. A., Assi, S., Unadkat, V. A., Patel, B., ... & Wardle, H. (2020). Poverty, inequality and COVID-19: the forgotten vulnerable. *Public health*, *183*, 110.
18. Abedi, V., Olulana, O., Avula, V., Chaudhary, D., Khan, A., Shahjouei, S., ... & Zand, R. (2021). Racial, economic, and health inequality and COVID-19 infection in the United States. *Journal of racial and ethnic health disparities*, *8*(3), 732-742.
19. Ali, S., Asaria, M., & Stranges, S. (2020). COVID-19 and inequality: are we all in this together?. *Canadian journal of public health*, *111*(3), 415-416.
20. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/908434/Disparities_in_the_risk_and_outcomes_of_COVID_August_2020_update.pdf
21. Ibid.
22. Sattar N, Hob F.K., et al. (2020). BMI and future risk for COVID-19 infection and death across sex, age and ethnicity: Preliminary findings from UK biobank. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* *14(5)*, 1149-1151

MEDICAL

Sharing data and expertise for the validation and updating of covid-19 related prediction models is urgently needed.[23] The issues concerning bias should be dealt with by making researchers adhere to reporting guidelines (i.e. TRIPOD or MINIMAR) to improve reporting and guide modelling choices, and by assessing risk of bias (e.g. by using PROBAST). At policy level, policy makers, public health officials, and AI developers should come together with affected stakeholders to figure out how to include vulnerable populations in policies, initiatives, and innovations.[24]

3.10 COVID-19 Diagnosis with AI

Viral nucleic acid testing and chest computed tomography imaging are standard methods for diagnosing Covid-19 but are time-consuming.[25] Thus, researchers have been developing diagnostic models to predict the presence and severity of covid-19 in patients in a more efficient manner.

These models are developed using single country data or international (combined) data.[26] Because pneumonia-signs on lung CT scans is one of the most common manifestations of Covid-19, several diagnostic applications focus on the use of image-recognition software to accelerate the reading of lung X-rays and CT scans.[27] There are more than 100 publications in MedRxiv and bioRxiv dedicated to this medical application.[28] However, there are other models being proposed that do not use images to diagnose the presence or severity of the virus. Some models work by using predictors such as vital signs (*e.g.* temperature, heart rate, respiratory rate, oxygen saturation, blood pressure), or flu-like signs and symptoms (*e.g.* shiver, fatigue), among others. Multiple different AI-based diagnostic tools have been proposed during the course of the pandemic. Yet, most of them have not been implemented at a larger scale but only in the context of small trials .

The COVID-PRECISE group reviewed and appraised the **technical robustness and efficacy** of published and preprint reports of 118 diagnostics models. This review showed that most models had a high or unclear risk of bias due to various reasons. Some models used inappropriate data sources, or other general research malpractices resulting in a high risk of bias. A high risk of bias implies that the reported accuracy of these models is too optimistic, thus the performance of these models in new cases will probably be worse than that reported by the researchers. Most models showed a lack of transparency and reproducibility due to a lack of descriptions of model specifications and subsequent estimations.[29]

In principle, this type of AI use can have a positive **impact on citizens and society** if developed responsibly. If however the methodologies are inadequate and the robustness is insufficient (as appears to be the case with many AI driven diagnostics models) the impact on citizens and society is negative, as it can result in incorrect, insufficient or lack of medical treatment. Given that diagnostics tools are used mainly for medical triage, high biases in these prediction tools can easily create space for unfair treatment among patients. Using unrepresentative data during the development of these models (as reviewed by the PRECISE group) can make the unrepresented population group be at a disadvantage when it comes to diagnosing them. In times of corona, this can lead to fatal consequences. Thus, a high or unclear risk of bias negatively impacts the right of citizens to be treated in a fair manner and potentially their health and life.

In January 2021, the FDA summarised its approach to a pre-market review of AI-driven software used for the diagnosis and treatment of disease. However, the PRECISE-group showed that there is a high/unclear degree of bias that is hidden behind claims of high accuracy claims in most studies. Such claims could allow models to be accepted through the pre-market review of AI-driven software. Therefore, there is an urgent need for researchers to administer proper **governance** adhere to reporting and **accountability** guidelines and to the use of appropriate calibration, and validations of their models as part of existing methodological guidance for prediction modelling studies.

3.11 COVID-19 Prognosis with AI

Due to the overcrowding of hospitals during the covid-19 pandemic, researchers have put efforts in developing AI models that predict the progression of the disease, with the aim of aiding with triage in times where resources are scarce. Most researched prognostic models aim to estimate mortality risk and progression to a severe or critical state of the disease. Other models, however, also aim to predict recovery, length of hospital stay, intensive care admission, intubation, the duration of mechanical ventilation, acute respiratory distress syndrome, cardiac injury or thrombotic complication. The most frequently used categories of prognostic indicators include age, comorbidities, vital signs, image features, sex, lymphocyte count, and C reactive protein. The models differ in prediction horizons (between one and 37 days).

---

23. https://www.covprecise.org/wp-content/uploads/2020/05/2021-Wynants-et-al-Update-4-BMJ-LR-prediction-models.pdf
24. Leslie, D. (2020). Tackling COVID-19 through responsible AI innovation: Five steps in the right direction. *Harvard Data Science Review (2020)*.
25. https://www.covprecise.org/wp-content/uploads/2020/05/2021-Wynants-et-al-Update-4-BMJ-LR-prediction-models.pdf
26. Ibid.
27. https://www.aaas.org/sites/default/files/2021-05/AIandCOVID19_2021_FINAL.pdf
28. Ibid.
29. https://www.covprecise.org/wp-content/uploads/2020/05/2021-Wynants-et-al-Update-4-BMJ-LR-prediction-models.pdf

The COVID-PRECISE group reviewed and appraised the **technical robustness and efficacy** of published and preprint reports of 107 prognostic models. The review showed that most models had a high or unclear risk of bias due to various reasons. Some suffered from dichotomization of predictors, or inappropriate inclusions/exclusions of study participants, leading to inappropriate sample sizes.[30]

Although the impact of accurately developing prognostic tools that could save lives would in principle have a positive impact on society, the state of research suggests that most current models suffer from a high degree of bias. The impact of using biased models can cause more harm than good, thus having a negative impact on **citizens and society**.

Similar to research on AI for diagnostics or identification of population vulnerability, the PRECISE group shows that there is an urgent need for researchers to adhere to reporting guidelines and be assessed against bias with tools such as to ensure that models are appropriately calibrated, validated, and using good modelling practices. Additionally, many reports did not report relevant information clearly.[31]

3.12 Cough Detection

As the name suggests Cough Detection applications are employed for the detection of COVID-19 through the sound of coughs. Many different smartphone apps are available, but they vary in their aims. Some aim to collect data for research purposes, while others aim to provide an instant diagnosis on COVID-19. For the latter, however, applications do not aim to substitute or replace formal COVID-19 tests or a medical examination. Examples of these applications include: *Cambridge sounds* app, *Breathe for science NYU* app, *Coughvid*, and *Voicemed* among others.

The level of **technical robustness and efficacy** of these applications is unclear. Many claim to be very accurate with percentages of more than 95% in detecting COVID-19 in cough samples, even of asymptomatic patients. Yet, many programs and applications provide no clinical evidence of their efficacy. For most applications, the data (recorded coughs, voice samples, etc.) were contributed by the users themselves, which could have hindered acquiring a diverse sample demographic group. If that is the case, the model could provide inaccurate results for people whose characteristics are underrepresented in the training data. Given that no verification and external validation procedures were in place, and in many cases, no medical professionals seem to have been involved in the process of collecting, categorising and verifying the data, more research in this area is required to assess the actual efficacy of using AI for COVID-19 detection through coughs.

Cough detection in principle has a ‘neutral’ **impact on citizens and society**. On one hand, clinically validated and used in clinical settings it could provide a low-cost early indicator of COVID-19 infections, which could help in triage. On the other hand, if these systems are widely used outside clinical settings in an unregulated manner, they could violate privacy regulations in relation to the use of biometric data. Furthermore, if these applications were to become mandatory or ‘hidden’ from the public eye, they could severely impact human agency, general privacy and physical autonomy.

For now, processing of biometric data conducted by these AI cough detection applications in the context of COVID-19 through the voluntary provision of coughs and voice samples follows the requirement for a legal basis for processing prescribed in the GDPR provided that the individual has given their explicit and free consent and the data is not used for other purposes.

However, risks may arise in the case of compulsory use of AI applications for cough detection for example by employees, as well as in the case of wide but covert use in public spaces. Especially for the latter, there is no **governance** structure in place.

3.13 Vaccine Hesitancy Chatbot

It has been argued that vaccines are our key in the fight against the pandemic. However, this only works when enough people are willing to receive it. The fast development of the vaccines has made many people doubtful of taking it.

---

30. Ibid.
31. Ibid.

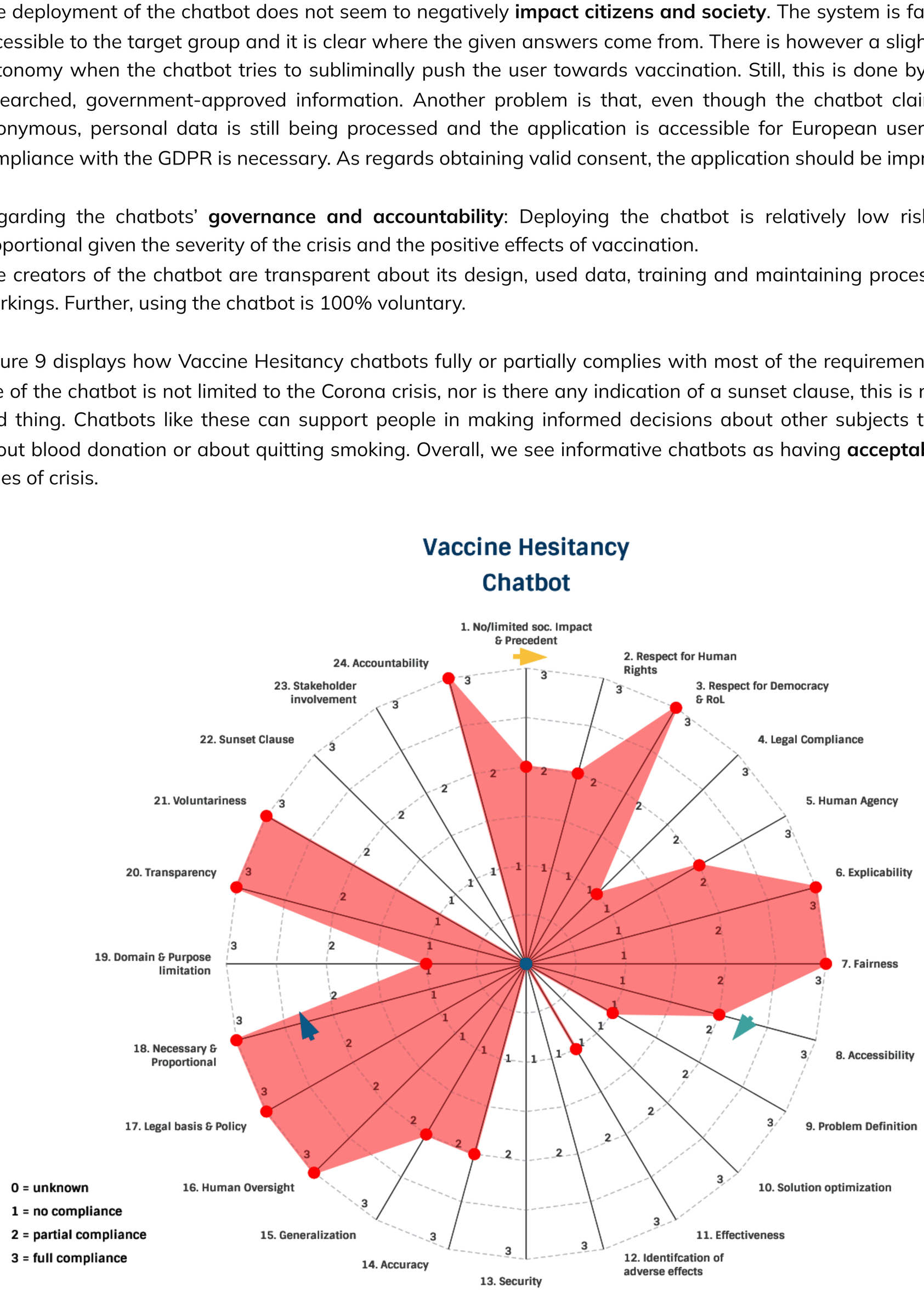
To combat vaccine hesitancy and misinformation researchers of the John Hopkins University and IBM have developed an AI-based chatbot.[32] The chatbot, named Vira, was inspired by a Japanese version that managed to increase the vaccine acceptance rate from 59% to 80%. Vira is supposed to help young people make more informed-decisions about taking or waiving a vaccination. It understands the main concerns surrounding vaccines and responds in a publicly acceptable, pro-vaccine way with answers written by experts on vaccinations. The application continually learns from new conversations, detecting emerging concerns and learning new ways in which existing ones can be expressed.

Even though chatbots are a possible solution to the vaccine hesitancy problem, their **effectiveness** has to be researched more. The chatbot does seem to work well and has a mechanism in place in case it malfunctions, making it more **robust**. However, there is no indication whether the risk of adverse effects has been researched, or whether any measures to secure the system are in place.

The deployment of the chatbot does not seem to negatively **impact citizens and society**. The system is fair, non-intrusive, accessible to the target group and it is clear where the given answers come from. There is however a slight risk for human autonomy when the chatbot tries to subliminally push the user towards vaccination. Still, this is done by providing well-researched, government-approved information. Another problem is that, even though the chatbot claims to be 100% anonymous, personal data is still being processed and the application is accessible for European users, meaning that compliance with the GDPR is necessary. As regards obtaining valid consent, the application should be improved.

Regarding the chatbots' **governance and accountability**: Deploying the chatbot is relatively low risk and therefore proportional given the severity of the crisis and the positive effects of vaccination.
The creators of the chatbot are transparent about its design, used data, training and maintaining processes, and overall workings. Further, using the chatbot is 100% voluntary.

Figure 9 displays how Vaccine Hesitancy chatbots fully or partially complies with most of the requirements. Although the use of the chatbot is not limited to the Corona crisis, nor is there any indication of a sunset clause, this is not necessarily a bad thing. Chatbots like these can support people in making informed decisions about other subjects too, for example, about blood donation or about quitting smoking. Overall, we see informative chatbots as having **acceptable trade-offs** in times of crisis.



32. https://vaxchat.org/

4

DISCUSSION & CONCLUSION

The above evaluations have shown that most AI applications developed and deployed during the Corona crisis have a negative impact on citizens and society, low technological robustness and efficacy and lack proper governance and accountability measures. Evaluation of these systems against the Framework for Responsible AI in Times of Corona has shown that the lack of compliance with the Framework's requirements is most apparent with surveillance applications such as face mask detection, thermal covid risk detection, movement tracking, online proctoring and home office monitoring. Moreover, with the exception of movement tracking applications, many applications suffer from a lack of technological robustness and efficiency, which is worrisome, especially when it comes to AI used for medical diagnosis and prognosis. The need for technological robustness appears to be often overlooked, probably due to "desperate times" causing public and private actors to take "desperate measures". Although most societal AI applications have an impact on human rights like privacy and physical autonomy, we believe that in cases where technical robustness and efficacy is high, the question of whether an acceptable trade-off between public health and fundamental rights existed, also comes down to a cultural context, and how much individual rights are valued over the public good or vice versa. Lastly, the only application that has shown to have the potential to positively impact citizens and society are AI applications used for vaccine discovery, cough detection for research purposes and informative chatbots. However, our findings also suggest that more research is needed to sufficiently prove the efficacy and robustness of the first two applications.

We are aware that results may be subject to change when looking at different use cases and technological developments. Nevertheless, our results show that there is an urgent need for more scrutiny and stricter measures that prevent the development and deployment of unsafe and irresponsible AI that can have a serious negative and even harmful impact.

The testing process of the Framework for Responsible AI in times of Corona has shown that it is well suited to execute a 'quickscan' when deciding on whether to develop, deploy or use an AI system to tackle an immediate challenge. While the Framework might look extensive at first, the equal spread of the requirements over the three main areas (society, technology and governance) provides a clear structure. The requirements themselves are clear and granular enough to move from overarching principles to practical application. The testing process also showed that for some use cases, not all requirements were relevant, however, skipping those requirements did not make the Framework less useful. Nevertheless, as a next step, we aim to develop adaptations to the Framework for specific AI uses or domains, such as healthcare.

Visualising the Framework through scores (0 - 3) and spider graphs showed to be useful to get a quick overview of the overall 'compliance' of the AI application as well as an idea of how the application scored per area.

ALLAI.

ABOUT ALLAI

ALLAI is an independent organisation that aims to foster, promote and achieve the responsible development, deployment and use of AI.

ALLAI's mission is to take a holistic approach to AI, taking into account all impact domains such as economics, ethics, privacy, laws, safety, labour, education, etc. ALLAI aims to involve all stakeholders in its mission: policy-makers, industry, social partners, consumers, NGOs, educational and care instructions, academics from various disciplines.

ALLAI was founded by the three Dutch members of the High Level Expert Group on AI, Catelijne Muller, LLM, Prof. Virginia Dignum and Prof. Aimee van Wynsberghe.

ALLAI refers to Stichting ALLAI Nederland, a foundation under Dutch Law. No entity or person connected to ALLAI, including its Board Members, Advisory Board Members, employees, experts, volunteers and agents, is responsible or liable for any direct or indirect loss or damage suffered by any person or entity relying wholly or partially on this communication.

CONTACT

ALLAI
Prinseneiland 23A
1013 LL Amsterdam
The Netherlands

www.allai.nl
@ALLAI_EU
welkom@allai.nl

© ALLAI 2022