Draft framework for responsible AI in times of Corona

Catelijne Muller, Virginia Dignum, Noah Schöppl

January 8, 2021

This is a publication by ALLAI, Foundation under Dutch Law, Prinseneiland 23A 1013 LL Amsterdam. This publication is part of the ALLAI Project 'Responsible AI and Corona'.

Abstract

In efforts to tackle the Corona-crisis, many public and private parties are considering to deploy or are already deploying AI-driven applications, both for medical as well as for societal uses. There is an increasing concern about the impact of these applications on fundamental rights, ethical principles and societal values, as well as their efficacy of controlling this pandemic. In this paper, we present a first outline of evaluation criteria to assess the efficacy and the legal, ethical and societal impact of AI-applications in times of Corona and requirements for their responsible use. This framework will be piloted with a selected group of organisations that are developing or deploying AI to tackle the various challenges of the Corona crisis.

1. Introduction

In efforts to tackle the Corona-crisis, many public and private parties are considering to deploy or are already deploying 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. There is an increasing concern about the implications of these applications on fundamental rights, ethical principles and societal values. Many of these applications are deployed through a fast-tracked decision cycle often driven by 'techno-solutionism'. 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 is 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.

This paper contains a first outline of evaluation criteria 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. These criteria will help organisations perform a 'quickscan' of the AI-application they want to develop, procure or deploy to tackle the challenges of this crisis. The quickscan 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.

2. Framework for Responsible AI in times of Corona (Quickscan)

The proposed evaluation framework is based on the Ethics Guidelines for Trustworthy AI of the High Level

Expert Group on AI[26] and on the ideas proposed in the paper "A socio-technical framework for digital contact tracing" by R. Vinueassa et al.[27]

It contains 24 criteria divided into 3 categories: (i) Technology; (ii) Impact on Citizens and Society; and (iii) Governance. Each criterion is measured on a scale from 0 - 3, where '0' is used when there is no information available and compliance with the requirement cannot be determined ('unknown', only reflected in *Figure 1*), 1 stands for 'non compliance', 2 for 'partial compliance' and 3 for 'compliance'. To acknowledge the context and urgency of this crisis, this framework aims to facilitate the timely and pragmatic application of responsible AI due diligence processes. It does not aim to provide a full and thorough assessment of the AI-application's technical robustness, societal impact and ethical and legal compliance, but rather an indication of what elements need to be assessed and what boundaries could be set in order to deploy AI responsibly. The use of the framework needs thus be complemented with deeper evaluation by experts from diverse backgrounds, and by users. Moreover, evidence (e.g. in the form of documentation) for the evaluation value should be openly available for inspection. We hope that this will help decision makers in their decision making process regarding the deployment of the AI-application.

It should be noted that these criteria are not meant to be used as a 'checklist'. The framework does not guarantee completeness (there may be more relevant criteria to take into account that are not included in such quickscan), and moreover the categories are interlinked and (can) serve as corresponding vessels. For example, an application could have a high level of efficacy but also have strong negative unintended consequences. The right governance (for example a sunset clause) might tip the balance in favor of using the application (in this case for a limited period of time). In contrast, an application with a low efficacy but a low impact level could nevertheless be interesting to deploy for research purposes.

In the following, we shortly introduce the criteria defined for each category, together with the meaning of their values.

2.1 Impact on citizens and society

1. **Societal impact, risk of undesired precedent**: The application has no identified undesirable impact on societal or human behaviour or adverse psychological effects and sets no dangerous precedent for the future (3). Some negative effects being identified or (2) or major negative effects being identified (1), is not adequate.
2. **Respecting human rights**: A comprehensive human rights impact assessment (HRIA) is undertaken prior to the system's development or deployment (3). Only some human rights risks having been considered (2) or multiple human rights being violated, the absence of an HRIA or the absence of a 'whistleblower mechanism', is inadequate (1).
3. **Respecting democracy and the rule of law**: The AI-system respects an open social and political discourse, free access to correct information, non-segregation and poses no risk of systemic failure or disruption (3). Only partially respecting/advancing democracy or the rule of law (2), or failure to respect or consider democracy and the rule of law (1), are not adequate.
4. **Compliance with existing legislation**: Depending on the application, this includes compliance with the GDPR, administrative laws, safety legislation, consumer legislation, worker legislation (a.o.) (3). Only partially in line with legislation (2), or not in line with legislation (1), are not adequate.
5. **Human agency**: The overall principle of user autonomy is central to the AI-system's functionality and use (3). Unclarity on or absence of this compliance with this principle (0) is inadequate.
6. **Explicability**: Explicability involves the right to explainability of outcomes of the system, especially when there is significant impact on the individual (3). Unclear or unsufficient explanations (2) or no explicability at all (1), is not adequate.

7. **Fairness**: The AI-application avoids unfair bias towards individuals or groups (3). Unclear measures to avoid this (2) or the lack of a plan to address this (1) is not adequate.
8. **Accessibility**: The AI-system can be used and/or is accessible for all, irrespective of age, demographic, disability, language, digital literacy and financial capacity (3). Addressing this only partially (2) or not at all (1) is not adequate.

2.2 Technology

9. **Problem definition**: There is a clear definition, understanding and description of the problem the AI-application aims to tackle (3). There is an unclear or partial problem definition (2). There is no clear problem definition (1).
10. **Solution optimization**: Alternative, less invasive solutions have been identified as less effective than the AI-application to solve the identified problem (3). Alternative solutions being as efficient as the AI-solution (2) leads to partial compliance. Alternative solutions being more efficient than the AI-solution, or not being identified at all leads to non-compliance with this requirement (1).
11. **Effectiveness**: The AI-application solves or significantly contributes to a solution for the problem (2). The AI-application only partially contributes to the solution (1). The AI application makes no contribution to the solution (0).
12. **Identification of the risk of adverse effects**: Potential adverse effects of the AI application in other areas are identified. These include effects on people, society and the economy: Adverse effects have been identified (3). Limited identification of adverse effects (2). Identification of adverse effects has not taken place (1).
13. **Security**: The AI system is protected against exploitations by adversaries, incl. dual use (3). Insufficient security processes or lack of policy to address these (1) are inadequate.
14. **Accuracy**: The AI system makes correct judgements based on data or models (3). Mitigation of the risk of inaccurate predictions and an indication of the inaccuracy percentage (2) or a lack of insight in the accuracy (0) is not adequate.
15. **Generalization**: The AI system exhibits reliable behaviour when used in a new situation (3). No generalization (0) is not adequate.
16. **Human oversight**: Human intervention and discretion during the entire operation of the AI-system is possible (3). Unclear measures to ensure this (2) or no possibility of human intervention and discretion (1) are not adequate.

2.3 Governance and accountability

17. **Legal basis and policy framework**: There exists a clear legal basis and policy framework for the use of the AI-system (3). Only partial legal basis or governmental policy (1) or no legal basis/policy (0) is inadequate.
18. **Necessity and proportionality**: The use of the AI system is necessary and proportional (3). Compliance with only one of these requirements (2) or no compliance at all (1) is inadequate.
19. **Domain and purpose limitation**: The goal and domain of the AI-application are clearly set and use of the application for other purposes or in other domains is strictly prohibited (3), or a clear policy stating the additional uses or domains is in place (2). Absence of domain and purpose limitation is inadequate (1).
20. **Transparency**: There is open and direct communication on the workings and the use of the AI-system, incl. open data governance, if necessary involving public education (3). Intermediate or private (2) is less adequate. Absence of communication on the workings, use or data governance (1) is not adequate
21. **Voluntary Use/Submission**: Using or being subjected to the AI application is voluntary (3). People cannot be denied access because of an obligation to use or be subjected to the AI application (1).
22. **Sunset clause**: The use of the AI-system is temporary, and a clear end date is set as well as a dismantling process (3). Unclear (2) or no sunset process (1) is inadequate.

23. **Stakeholders:** Relevant stakeholders have been involved in the development/deployment process (3). Only partial stakeholder involvement (2) or no involvement at all (1) is not adequate.
24. **Accountability**: There is clear 'ownership' of the application and proper documentation available on the workings and use of the AI system (3). Insufficient, partial (2) or no clear 'ownership' or documentation (1) is not adequate.

2.4 Tensions, trade-offs and boundaries in times of crisis

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 Ethics Guidelines of the High Level Expert Group on AI[1] provide guidance for such a 'balancing exercise':

- the benefits of AI must outweigh the individual and collective risks;
- particular attention should be paid to vulnerable groups such as children;
- the fact that AI carries risks that are difficult to predict should be recognized;
- adequate and proportional measures should be taken to prevent or reduce those risks;
- there may be situations where there is no ethically acceptable trade-off and the application should not be used;

The current extraordinary circumstances 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.

3. Example of application of the framework: 'AI-driven mask wearing detection at Châtelet-Les Halles metro station'

AI technologies are being used or tested to detect public adherence to Corona measures such as mandatory mask wearing. The French RATP cut short a test of mask-detection at its Paris metro station Châtelet-Les Halles after criticism, among others by the French privacy watchdog CNIL. CNIL predominantly looked at the impact of the technology on human rights and the GDPR and called for vigilance in using these kinds of surveillance technologies. We have tested the technology against the broader requirements of the framework in this paper. Apart from lack of clarity on a large number of requirements, because of the lack of publicly accessible information, the application never scored above the non-compliance threshold (1). See *Figure 1*

1

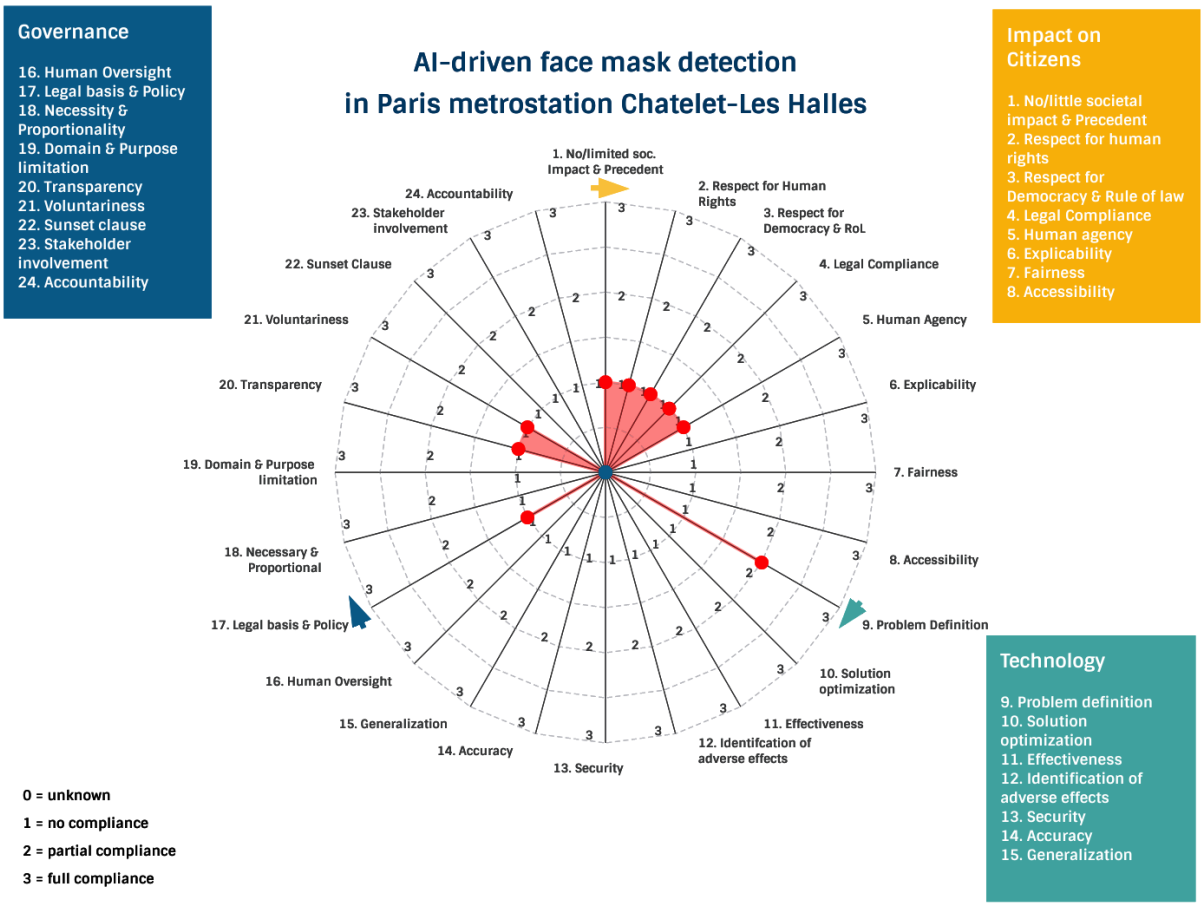

Figure 1. Visualisation of compliance with framework of AI-driven face mask detection cameras at Paris metro station Châtelet-Les Halles

3.1 Impact on citizens and society

The **societal impact** of this AI system can be considered high, as well as the risk of **undesirable precedent** leading to the lowest score (1). The mask detection system at metro station Chatelet-Les Halles does not appear to be standing on its own, but appeared to be part of a larger AI laboratory at RATP, which was supposed to start its operations mid 2020. This laboratory aimed to test different AI surveillance systems at the metro station with the objective to learn to detect "problematic situations" and "abnormal" behaviors.[2] In a note entitled *Security in the Age of Artificial Intelligence,* the Paris Region Institute (IPR) indicates that the RATP, deploying over 50.000 security cameras, is regularly requested by manufacturers to "participate in research programs aimed at improving the technicality of algorithms".[3]

There is an **impact on a number of human rights**, starting with the right to privacy. It should be noted that the impact of AI on our privacy goes well beyond our personal data. Art. 8 of the ECHR encompasses the protection of a wide range of elements of our private life, that can be grouped into three broad categories namely: (i) a person's (general) privacy, (ii) a person's physical, psychological or moral integrity and (iii) a person's identity and autonomy. The mask detection system involves the capture of personal (biometric) data (our faces) , but it also affects our right to a private life and autonomy in such a way that it creates a situation where we are (constantly) being watched and followed. As a

[2] https://www.ladn.eu/tech-a-suivre/ratp-chatelet-halle-laboratoire-intelligence-artificielle-surveillance/

[3] https://www.institutparisregion.fr/fileadmin/NewEtudes/000pack2/Etude_2310/NR_833_web.pdf

psychological 'chilling' effect, people might feel inclined to adapt their behaviour to a certain norm such as avoiding public transport. Thus, the application also scores lowest as regards respect for human rights (1).

As regards the impact on **democracy and the rule of law**, CNIL argued that "(...) anonymity in the public space is an essential dimension for the exercise of these freedoms and the capture of the image of people in these spaces undoubtedly carries risks for their fundamental rights and freedoms." CNIL also argued that "(...) Their uncontrolled development presents the risk of generalizing a feeling of surveillance among citizens, of creating a phenomenon of habituation and trivialization of intrusive technologies, and of generating increased surveillance, liable to undermine the proper functioning of our democratic society." The application scores (1) on the requirement for respect for democracy and the rule of law.

CNIL also indicated that the application/use case did **not comply with the law**, i.e. the right to object against the collection of a person's image (art. 21 GDPR). The suggested solution where people could 'shake their head' if they did not want to be captured on camera, was deemed insufficient and impractical, causing the application to score (1) for legal compliance. CNIL also argues that: "If the right of opposition cannot be applied in practice, the devices concerned must be specifically authorized by a specific legal framework provided for either by the European Union or by French law." This means that the application also does not comply with our no. 17 requirement of **a legal basis and governance framework** giving it a (1) score as well.

As regards respect for **human agency**, making sure people are not subjected to AI systems that unconsciously influence behaviour or make decisions about them, without the possibility to avoid or object to these, the AI-system also scores the lowest (1). Subjection to the AI-system could only be avoided by not using the Chatelet-Les Halles metro station, thus limiting people in their freedom to move about. Moreover, the system could have a behavioural 'chilling-effect' on people who could feel 'watched' all the time without knowing by whom.

No publicly available information could be found on requirements of **explicability** of the application, **fairness** in the sense of freedom from bias and equal treatment of all that are subjected to it, and **accessibility**. The application thus scores (0) for requirements 6 through 8.

3.2 Technology

As regards to the '**problem definition**' and solution optimization (requirements 9 and 10), we looked at the function of the application: 'detection of people wearing or not wearing a mask in public spaces' and its goals described by the developer: 'health security management for staff and the public; distribution of masks adapted to needs and promotion of masks in the public space.'

For the trial at Chatelet-Les Halles we gathered from the media that RATP wanted to test 6 camera's to "make it possible to have, in real time, an estimate of the number of travelers who comply with health instructions." How this estimation would then be used, e.g. to determine whether mask wearing should be promoted more or to inform policy making or for law enforcement, is unclear. We find the problem definition inadequate but not completely lacking, thus giving it a score of (2). We have not found any information on whether any **other solutions** were considered (0).

For **effectiveness**, we looked at the extent to which the application solved or contributed to solving the (partially) identified problem. Because the test was abandoned prematurely, we could not score this requirement (0).

We have not found any information on whether possible **adverse effects** such as chilling effects on citizens, undesired precedent, etc. were identified, giving it a (0) score.

There is no information available on the resilience of the application against cyber attacks or misuse, giving it a (0) score for **security**.

As regards **accuracy and generalization**, a video of the workings of the application shows 4 white men directly facing the camera, with and without proper mask wearing, where mask wearing is correctly detected.

The developers refer to research[4] on how to address issues with face alignment and head pose, gaze estimation and image completion, but it is not clear whether the methodologies presented and/or proposed in this research have been applied to the application.

We thus have insufficient information to determine the effectiveness of the application in actual situations and circumstances, e.g. where people do not directly face the camera, pass at different speeds, wear scarfs or hoods, have a darker skin color, are poorly lit, are in a crowd, etc. The application thus scores (0) on accuracy and generalization.

No publicly available information could be found on requirements of **human oversight** over the application, giving it a (0) score.

3.3 Governance

CNIL considered that these types of applications need a legal basis under the GDPR. We have not found any indication of compliance with this requirement for this system. The developers claim that the application is in line with the GDPR as it does not store any data.

[4] "Deep Entwined Learning Head Pose and Face Alignment Inside an Attentional Cascade with Doubly-Conditional fusion". By Arnaud Dapogny 1,2, Kevin Bailly1,3 and Matthieu Cord 2,1; also "DeCaFA: Deep Convolutional Cascade for Face Alignment In The Wild". By Arnaud Dapogny1, 2, Kevin Bailly2, 3, and Matthieu Cord1; also "Tree-gated Deep Mixture-of-Experts For Pose-robust Face Alignment". By Estèphe Arnaud1, Arnaud Dapogny2 and Kévin Bailly1, 2; also "The Missing Data Encoder: Cross-Channel Image Completion with Hide-And-Seek Adversarial Network**"** by A. Dapogny, M. Cord and P. Perez, AAAI 2020

We however also argue that these types of applications, that are administered in times of crisis, need a **legal basis** beyond the GDPR. The invasiveness when it comes to human rights and freedoms and our democracy (as also identified by CNIL) warrants that if at all necessary and proportionate, these types of applications should only be allowed after democratic scrutiny. This could mean under emergency regulation for example.

As far as we could determine, France is only considering new legislation to create a framework for wide-ranging tests of facial recognition at this point. Score (1).

We could not determine the **necessity & proportionality** of the system, primarily due to the absence of an adequate problem definition, giving it a (0) score. The same goes for the need for **domain and purpose limitation**, i.e. limiting the use of the system and its in- and outputs for the specific goal of tackling the defined problem only, leading to a (0) score.

We did not find any information on whether the public was made aware of the fact that they were subjected to a facial recognition system that detected the proper mask use at the metro station in an open and **transparent** manner. Score (0).

Consent is one of the conditions in the GDPR that could justify the processing of (special categories of) personal data and could thus guarantee **voluntary use of, or submission to** the system. We argue however that because of the invasiveness of this system, the fact that it scores low on all requirements, especially on respect for human rights and democracy, voluntariness should be unconditional and thus also be adhered to, if for example one of the other conditions of the GDPR is met. Practically, voluntary submission to the system seems only possible by avoiding the metro station all together.

The developers indicated that people could 'shake their head' if they wanted to express opposition to the use of the system. CNIL however found that this "does not appear satisfactory from the point of view of the protection of the interest of people. This solution is impractical in practice and difficult to generalize.

It also forces individuals to publicly display their opposition and places too great a burden on the person, all the more so if such devices multiply. In principle, such modalities does not ensure the effectiveness of the right to object must therefore be considered as non-compliant with the provisions of the GDPR." The system thus scores the lowest on voluntariness (1).

Information on whether a **sunset clause** for the use of the system was considered, whether **stakeholders** were involved and if an **accountability** process was in place was not available. Score (0) for requirements 22 through 24.

4. Conclusion

We conclude that the face mask detection system of the RATP shows an overall low score, but one of the main issues with this particular use case is the lack of publicly available information on the system. It shows that without adherence to the requirements of transparency and accountability, the AI application cannot meet many of the other requirements either.

Acknowledgements

Contributor: Dr. Jan van Gemert, Head of the Computer vision lab, Faculty of Electrical Engineering, Mathematics and Computer Science Delft University of Technology. ALLAI received funding from SIDNFonds for the Project Responsible AI & Corona.