# A Scenario-Based Approach to the Design and Use of Ethical AI Models in Managing a Health Pandemic

Georg Groh[1], Dirk Brand[2], Johan van der Merwe[2,3], McElory Hoffmann[2,3], Tobias Eder[1], Edoardo Mosca[1], Ben Herbst[2,3], Aldi Topalli[1], Miriam Anschütz[1], Fatos Morina[1], Yifeng Dong[1], Maria Angeles Lopez Forner[1], Soh-Yee Lee[1], Tuhin Ghosh[1], David Elias Drothler[1], Timo Fischer[1], Joshua Arnold[1], Min Wu[1], and Grigor Bezirganyan[1]

[1]Technical University Munich
[2]Stellenbosch University
[3]Praelexis AI

December 2021

## 1 Introduction

### 1.1 Project description

Efficient acquisition and processing of information is a key factor in controlling pandemics. Relevant elements of the health status (e.g. infection state) of individuals or groups of individuals, mobility patterns of citizens, or information on social contacts and social networks are examples of such information. Information processing for decision making may encompass machine learning models for the analysis of the dynamics of an epidemic using and linking such data. However, privacy interests and other ethical issues have to be carefully considered.

This inter-disciplinary research project aims at investigating approaches for information acquisition and machine learning information processing. The goal of the analysis is to provide an ethical basis for policy decisions in which AI models are used to develop and use viable combinations of measures.

### 1.2 Research Objectives

#### Socio-ethical issues

Health-related management decisions and responses to threats like pathogens and the resulting pandemics are governed strongly by the foundational principles of bodily integrity and human dignity. When a machine learning algorithm is involved in the decision-making process, aspects of privacy, fairness, accountability and interpretability become fundamentally entwined in both the deliberation process and the execution of health-related management and data policies. The research aimed to explore these dimensions as it relates to data governance in health-related settings and to devise an ethical framework describing the lenses to be held over healthcare and data management decisions. The expectations and rights of individuals, communities and institutions of governance were placed alongside the predictive and prescriptive value of algorithms in the deliberative process of formulating a response to a health-related issue. An ethical framework was applied to the different scenarios identified in the research.

#### Legal and governance issues

Technological developments must serve the people. It is therefore important to ensure that AI solutions are developed within a constitutional legal framework characterised by the rule of law and the protection of human rights [19]. Against this background this research project focused on the development of an appropriate legal and governance framework for the design and application of an AI model to manage health pandemics. Some of the critical questions considered in this process are: Which information gathering and processing measures are legal? What are the elements of a legal and governance framework for AI health management models? How could this be used in different jurisdictions in the world?

#### Machine Learning

Effective machine-learning approaches have to be developed and evaluated in each plausible data scenario, in view of monitoring and controlling the spatio-temporal and social dynamics of an epidemic as well as the different settings in which information can be linked or are available. One key aspect of such machine learning approaches is explainability [45] [40] [54], i.e. the ability of the models to justify their results and make them transparent to a human. These approaches must maximize predictive power and must be practically applicable (e.g. in scenarios where federated learning [66] is deemed the only appropriate approach), while addressing identified ethical and legal issues.

#### Key research questions

The key research questions are:

- What are the socio-ethical considerations when generating and processing personal data in health-related settings in terms of privacy, fairness and agency and what are the shortcomings of current approaches?
- What are the elements of a legal framework for AI-based epidemic management models? How could this be used in different jurisdictions in the world?
- What are possible machine learning strategies that could be applied in managing major health crises based on an ethical and legal framework acceptable within a constitutional democracy?

## 1.3 Research Methodology

The research methodology is a design science approach. We assume an epidemic scenario involving a communicable disease involving a pathogen. Data was generated by simulation, which involved the following:

### Individual, node-profile data

Medical data (e.g. infectious status) according to state-of-the-art epidemiological and medical models and according to testing strategies for epidemic cases (e.g. Covid-19, HIV, Ebola) [6].

### Social, edge-profile data

Human contacts as instantiations of dynamic social networks (generated by dynamic social network models [41][69] and spatio-temporal human mobility patterns (generated by state of the art mobility models [7] [63] in connection with node-profile data and epidemiologic models [32] [48] [23] [34].

In order to ensure validity of the generated data we analysed its compatibility with real world data, where available. With the simulated data, a number of plausible scenarios were built in terms of degrees of accuracy of the data (e.g. spatio-temporal resolution) and certain degrees of availability of the data. These degrees of availability might be due to various testing policies, certain data-gathering policies, or certain ethical concerns.

Using the data scenarios, we assessed and developed suitable machine learning and data-analytical approaches for momentary analysis of the situation (e.g. detection of infection hot-spots via clustering) and for the predictive analysis of the situation (via epidemiological models). We then assessed their feasibility and suitability from a socio-ethical, legal (governance), and political point of view. This assessment is used to refine the generation of scenarios and the machine learning and data-analytical approaches in the next cycle of the design science approach.

## 1.4 Clarification of concepts

While this report is primarily aimed at well-informed readers, it is envisaged that in view of the focus of this study, that it could be of interest to policy makers in government in different countries who might not be equipped with good knowledge of this topic. Some key concepts used in the research study and report are therefore clarified here.

**Artificial intelligence (AI)** could be defined as 'the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages' (Oxford Dictionaries, 2020). AI is increasingly being used in the field of health care in a variety of applications. The use of AI, including machine learning, is about automated decision making that can assist or replace human decision making. In the context of this research study the focus is on ethical AI models that can assist human decision making to manage health pandemics. The ethical considerations provide a value-base for the discussion of the legal concepts such as accountability and transparency and the application thereof in the use of AI in fighting pandemics. Technological developments must serve humanity and have a net positive effect. The design and use of AI should thus follow a human-centred approach. It is therefore important to ensure that AI solutions are developed within a constitutional legal framework characterised by the rule of law and the protection of human rights.

"**AI ethics** is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies" [37].

**Machine learning** can be defined as the science of getting computers to learn and act like humans do by using algorithms and learning from it, having the ability to learn without being explicitly programmed [43]. With machine learning computers learn from data and create solutions to complex problems, including predictions based on the knowledge it gained.

Dignum [20, p.3] argues that "AI represents a concerted effort to understand the complexity of human experience in terms of information processes. It deals not only with how to represent and use complex and incomplete information logically but also with questions of how to see (vision), move (robotics), communicate (natural language, speech) and learn (memory, reasoning, classification)." Within the broader field of AI, Machine Learning (ML) enables computers to sense, to comprehend and to act in such a way that they automatically improve by experience and correctly perform tasks in new settings (Praelexis, 2020). Usually, ML is classified into three kinds of algorithms:

1. **Supervised learning** is a type of machine learning algorithm that uses a known dataset (called the training dataset) to make predictions. The training dataset includes input data and target values. From it, the supervised learning algorithm seeks to build a model that can make predictions of the target values for a new dataset.

2. **Unsupervised learning** is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled targets. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data.

3. **Reinforcement learning** is a type of machine learning algorithm that uses a system of reward and punishment. A reinforcement learning algorithm, or agent, learns by interacting with its environment. The agent receives rewards by performing correctly and penalties for performing incorrectly. The agent learns without intervention from a human by maximizing its reward and minimizing its penalty. (Praelexis, 2020)

### 1.5 Context of research project

The outbreak of the Covid-19 pandemic in December 2019 has created a sense of urgency that clearly expedited the use of technology in the health industry, with its newfound focus on the benefits in pandemics. Pandemics do pose unique challenges that calls for bespoke and sometimes novel applications. Studies reviewing the application of AI in Covid-19 (cf. [35] [14]) have indicated the use of AI systems for dealing with the pandemic. AI and ML is making a crucial contribution to the treatment, medication, screening, prediction, forecasting, contact tracing, and drug or vaccine development process for the pandemic and reduce the human intervention in medical practice [35, p. 1].

# 2 Ethical AI and design for values

## 2.1 Ethical Principles applied to AI and ML Algorithms

### 2.1.1 Responding to the properties of AI and ML Systems

AI systems have "agency", which entails some degree of autonomy, adaptability and interactivity. Responsible AI aims to channel these attributes in ethically sound directions, doing so by defining responsibility, transparency, and accountability.

**Autonomous Systems requires Responsibility** Although AI systems can make decisions, select best next actions, et cetera, human responsibility cannot be replaced. Even if a system is designed for accountability and transparency, it is still an artefact or tool constructed by human intervention where a purpose is determined for the system. Even if the system can modify itself and learn from its context, it still performs according to the purpose determined by the human designer. While there is some discussion around assigning responsibility to AI systems itself, the options for responsibility really comes down to (i) the machine acts as intended and therefore the responsibility lies with the user, as is the case with any other tool; or (ii) the machine acts in an unexpected way due to error or malfunction, in which case the developers and manufacturers are liable [20, p.57]. Actions of a system because of learning cannot remove the liability from its designers and developers, as it remains a consequence of the algorithms they designed. This is why the design and implementation processes need to be closely monitored to ensure that the system is behaving according to the relevant ethical and societal principles.

**Adaptable Systems requires Transparency** AI systems learn from data and is trained on subsets of data in order to be able to interpret new data and predict new values. Not only do systems need to be able to account for actions and explain how decisions were reached, but they also need to be transparent, i.e. open for inspection and monitored for bias and drift. The need to deal with bias in data and algorithms and monitoring drift is easier said than done, because data inevitably show patterns and take features into account in decision-making. For example, it cannot be construed as bias if a loan is not granted to someone with a large propensity to default and low affordability. However, it will be wholly unethical to distinguish according to, say, gender or race.

Even if it may be illegal to use certain attributes in decision-making, such as gender, certain patterns are sometimes discovered by the Machine Learning algorithm, and the system can use them as a proxy (e.g. for gender), thus reinforcing bias. The aim of algorithmic transparency is to ensure that the machine will not be prejudiced and act on these biases in the data [20, p.60].

Besides bias, problems with data can include incompleteness, tampered data, and outdated data. Transparency also needs to deal with these complexities and provide openness and control over the whole design, training, productionization, and monitoring process. Models can malfunction when changes in the underlying relationship between input and output data, called "concept drift" in Machine Learning, are not detected and mitigated.

Handling bias and drift is all part of transparency in the management of ML models. It includes openness about data, design processes, functioning of the algorithms and a keen awareness of the participation and involvement of all relevant actors and stakeholders.

**Interactive Systems requires Accountability** AI systems interact with other systems and users. Accountability is the capability to give account of those interactions, that is to report and explain one's actions and decisions. The difference between responsibility and accountability is important. Responsibility refers to the duty to answer for one's role in actions and entail liability. It exists even before the action, and when an action is delegated (to another person or a system) the delegating person remains liable for the consequences of the action. Accountability refers to the ability to explain or report on one's role in events. It is only evident after the action is done. So, while the delegating person remains the "principal" responsible and liable, the agent (which can be an AI system) must be able to report and explain how tasks were executed [20, p.54].

The accountability is not only for a system to be able to explain why it took a certain course of action, but also to prove that a safe and sound design process were followed, incorporating the relevant values and ethical concerns. Accountability in AI systems therefore refers to "explainable AI". Explanation reduces the opaqueness of a system (the so-called "black box") and when things go wrong, the process can be audited, and logging systems can indicate where errors crept in (similar to the role of the flight recorders in aviation disasters). Accountability in AI systems are particularly important, given the fact that the system lacks moral agency and therefore need to "explain" how design processes and the results of algorithms incorporated ethical "reasoning". Thus the need for a "glass box" and not a "black box".

### 2.1.2 Lenses for Ethical Reasoning

Given the principles for responsible AI, the question is how to assess ethical dilemmas in order to reach a mature level of ethical reasoning and actionable decisions. The need for moral and ethical reasoning goes back to the dawn of humankind. Overarching normative ethical theories emerged over the course of time to settle ethical dilemmas. Usually, three main theories are mentioned, namely consequentialist, deontological and virtue theories, e.g. Boddington [10, p.8]. Vallor et al [60] expands this to five lenses or perspectives for ethical reasoning and apply this to technology best practices:

**The Rights Perspective** Also referred to as deontological ethics, this ethical lens focuses on moral rules, rights, principles and duties. It intends to be more universally applicable to the majority or all possible cases (as opposed to a situational ethics approach that tends more to the context of an ethical dilemma). Still, instances will invariably arise where principles are in conflict and minimizing the ethical violation and prioritizing the principles are inevitable. Principles may be general, like the "golden rule" ("do unto others as you would like them to do to you") or Kant's "categorical imperative" ("act only according to that maxim whereby you can, at the same time, will that it should become a universal law") or Kant's "formula of humanity" ("humans are required never to treat others merely as a means to an end, but always as ends in themselves"). Otherwise, lists of moral duties can also be cited (e.g., Ross's fidelity, reparation, gratitude, justice, beneficence, non-injury and self-improvement) [60, 2].

When applying the perspective to issues emerging in the interaction between humans and AI systems, principles like autonomy, dignity and transparency are important. People are to choose for themselves, be valued in themselves and be informed in order to grant consent. This does not mean that technology should empower users to do anything they want. The principle of autonomy needs, for example, to be balanced under certain conditions with "appropriately limited moral paternalism" [60, 2]. Questions that might direct ethical reasoning in considering ethical principles and duties include "How might the dignity and autonomy of each stakeholder be impacted by this system?" and "Are our choices of the sort that we could find universally acceptable?"

**The Justice or Fairness Perspective** This lens wants to ensure the appropriate distribution of benefits and burdens, taking into consideration ethically relevant distinctions among people (known as distributive justice). Elements such as need, contribution and the impacts of social structures on individuals and groups also need to be considered [60, 3]. Retributive justice (for example punishing criminal wrongdoing) and compensatory justice (compensating disadvantaged people) also forms part of the justice perspective.

Justice goes along with fairness and impartiality. It entails also the "veil of ignorance" (John Rawls) about characteristics over which people have no control (age, gender, race, et cetera). This would allow for a more egalitarian and fair society. Especially attentive to the relationships among people the justice perspective preserves dignity in terms of societal obligations. Concerns that need to be heeded, also in the assessment of ethical issues with AI systems, are equality, equity, fairness, diversity, inclusion, due process and consistency.

Important questions to be asked would include "How are the benefits and burdens of this system distributed among various stakeholders" and "Does the system exclude any vulnerable or marginalized groups from participation and opportunity?"

**The Utilitarian Perspective** Consequentialist theories claims that the right action is the one that brings about the best consequences. This is most commonly held as some form of "utilitarianism", which aims to bring about the greatest balance of happiness over unhappiness, or pleasure over pain, for the largest number of people [10, p.8]. The theory is attractive because in seems to promise optimal results, at least in theory. Because the consequences of technology spread out indefinitely in time (do we know, for example, what the net positive or negative effect of social media will be for humanity?) it often leads to senseless calculations. However, utilitarian ethics does prompt the consideration of consequences for more stakeholders than would otherwise be included if only a limited cost-benefit analysis for selfish gains would be done. Moral consequences go far beyond economic good and harm. They include not only physical, but psychological, emotional, cognitive, moral, institutional, environmental, and political well-being or injury, or degradation [60, p.5]. Considering comprehensive happiness, balancing the interests of all stakeholders and the calculation of all reasonably foreseeable effects of a system, would be a meaningful application of this ethical theory when assessing ethical matters with regards to AI systems. Relevant questions would include "How might future generations be affected by the effects of this system?" and "Will this system create more good than harm, and what types of good and harm?"

**The Common Good Perspective** While utilitarianism focuses on the benefit or harm to individuals and then sum those up to measure aggregated social impact, the common good perspective focuses on the impact of a practice on the well-being of communities and eventually humanity as functional wholes [60, p.7]. Beyond mere personal happiness, this perspective would want to enhance political and public well-being, sustainability, education and flourishing communities. Technology acceptable to a utilitarian because it makes most individuals happy, would not necessarily satisfy the common good ethicist if loss of community flourishing would occur in the process. These ethicists will therefore also guard the welfare of societal institutions like government, the justice system, education system or ecosystems.

Assessing the impact of technology on communities, relationships, institutions of governance, economic and other social institutions would be the application of the common good lens. Relevant questions may include “How might this technology benefit social institutions?” and “How might this technology do for the environment beyond human society, such as biodiversity and sustainability?”

**The Virtue Ethics Perspective** Virtue ethics focuses on the character of the ideal moral agent, describes the range of different virtues such an agent has, and claims that the right thing to do in any given situation is to do what the fully virtuous person would do [10, p.8]. This type of ethics recognized that any set of moral rules is incomplete and people with moral character will need to fill the gap with wise moral judgments. Most of moral life is not a closed problem where one optimal solution exists. It entails navigating a messy, open- ended and constantly shifting landscape [60, p.8].

Virtue ethics does guide us, though, by offering those traits that would, if it is constantly exhibited in actions and habits, promote virtual lives and societies. If honesty is a virtue, for example, then we need to think of habits of design in AI systems that prevent falsified data to serve as input for models. Virtue ethics is situational ethics which is very sensitive to context. Wisely distinguishing between the options for appropriate action according to highly developed moral perception, moral emotion and moral imagination lead to well-chosen courses of action for particular circumstances. An action that would be virtuous in one situation, may be foolish in another.

Relevant questions to ask would include “Would we want future generations of data scientists to use our current practice as an example to follow?” and “What habits of character will this design or system foster in users and other stakeholders?”

**Making Ethical Decisions** The Markkula Center for Applied Ethics [30] suggests, on the basis of the five lenses described above that a framework in the form of ten questions can be applied to facilitate ethical reasoning:

*Recognize an Ethical Issue*

1. Could this decision or situation be damaging to someone or to some group? Does this decision involve a choice between a good and bad alternative, or perhaps between two “goods” or between two “bads”?

2. Is this issue about more than what is legal or what is most efficient? If so, how?

*Get the Facts*

3. What are the relevant facts of the case? What facts are not known? Can I learn more about the situation? Do I know enough to make a decision?

4. What individuals and groups have an important stake in the outcome? Are some concerns more important? Why?

5. What are the options for acting? Have all the relevant persons and groups been consulted? Have I identified creative options?

*Evaluate Alternative Actions*

6. Evaluate the options by asking the following questions:
   - Which option will produce the most good and do the least harm? (The Utilitarian Approach)
   - Which option best respects the rights of all who have a stake? (The Rights Approach)
   - Which option treats people equally or proportionately? (The Justice Approach)
   - Which option best serves the community as a whole, not just some members? (The Common Good Approach)
   - Which option leads me to act as the sort of person I want to be? (The Virtue Approach)

*Make a Decision and Test It*

7. Considering all these approaches, which option best addresses the situation?
8. If I told someone I respect—or told a television audience—which option I have chosen, what would they say?

*Act and Reflect on the Outcome*

9. How can my decision be implemented with the greatest care and attention to the concerns of all stakeholders?
10. How did my decision turn out and what have I learned from this specific situation?

## 2.2 The Ethics of Algorithms

How we perceive and navigate our environments is increasingly mediated by algorithms which advise, if not prescribe and mandate our best next actions. In that sense, algorithms are inescapably value-laden [44, p.1]. The ethical gap between the intended purpose designed into an algorithm (which purpose in itself can also be dubious) and the implementation and outputs of these algorithms in reality can have severe consequences affecting individuals and societies. Mittelstadt [44, p.4ff] mapped the ethics of algorithms in terms of six types of ethical concerns raised by the algorithms. Three of these concerns are of an epistemic nature (how data is turned into evidence and outputs, namely inconclusive, inscrutable and misguided evidence) and two are of a normative nature (unfair outcomes and transformative effects). Potential failures involving multiple stakeholders complicate the question of who is responsible and accountable, and that leads to the last type of ethical concern which is traceability.

1. **Inconclusive evidence** refers to the nature of algorithms drawing conclusions from data with degrees of uncertainty and probability. Moreover, conclusions are often indicating the probability of correlations, but seldom the existence of a causal connection (causal modelling is another area of machine learning altogether and is currently being researched intensively). When generating "actionable insights", it has to be viewed responsibly and recognizing limitations. Inconclusive evidence can lead to unjustified actions [44, p.5], for example acting on individuals when insights generated with some probability of only

correlation may concerns populations as a whole, such as when an insurance premium is set for a sub-population but now has to be paid by each member.

2. **Inscrutable evidence** refers to the often-encountered issue that the connection between the data and the conclusion is not always accessible and intelligible. Not being able to interpret how many data-points are used by an ML algorithm to generate conclusions cause practical and principled limitations. Inscrutable evidence leads to opacity, obscuring the accessibility and comprehensibility of information. This becomes particularly troublesome when an imbalance of knowledge and decision-making power arise where data subjects are not privy to how algorithms take decisions on their data (e.g., in credit ratings).

3. **Misguided evidence** refers to the GIGO (Garbage-In-Garbage-Out) principle that output can never exceed input as far as data is concerned. Conclusions can only be as reliable as the data they are based on. Misguided evidence leads to bias, "freezing" the values of the developer into the code. An example of this is when ML algorithms are trained from human-labelled data training sets. The biases of the taggers will be reflected in the functioning of the eventual model.

4. **Unfair outcomes** focuses on the actions driven by the algorithms. An action can be discriminatory solely from its effect on a vulnerable group of people, even if made on the basis of well-founded evidence. Unfair outcomes lead to discrimination, of which profiling algorithms is a good example. Whereas bias is a dimension of the decision- making itself, discrimination describes the effects of the decision [44, p.8]. Even when algorithms are directed to disregard sensitive attributes, proxies can form and continue the discriminatory effects. Another example is personalization or dynamic pricing, which can segment a population or individualize offerings to such an extent that only some segments receive the opportunities or information, re-enforcing existing social inequalities [44, p.9].

5. **Transformative effects** refer to the ways in which seemingly neutral data- processing and data-driven decision-making can affect how we conceptualize our world in new ways, which can be questionable. Transformative effects lead to challenges for autonomy, for example when recommender systems nudge the behavior of data subjects and decision-makers by filtering information. While supporting decision-making, it may also be controlling decision-making, thereby manipulating data subjects and eroding individual autonomy. This is all the more problematic when the desired choice reflects third party interests above those of the individual. Also, filtering algorithms can create "echo chambers" where the diversity of information relayed to users decrease and decisional autonomy is thus impeded [44, p.9]. Transformative effects also may lead to challenges for informational privacy, meaning the capacity of individuals to control information about themselves and the effort required by third parties to obtain this information. Even when data are anonymized, profiling and segmentation cause linking of an individual to others within a dataset and judging or treating the individual on that basis.

6. **Traceability refers** to the complexity of detecting and tracing the cause of harm done by algorithmic activity. This is also relevant with regards to the assignment of responsibility and accountability. Good traceability leads to the acceptance and distribution of moral responsibility. How malfunctioning algorithms or unexpected outcomes will be handled in

terms of the apportionment of responsibility, remains a lively debate. The gap between the designer's control and algorithm's behavior creates an accountability gap wherein blame can potentially be assigned to several moral agents simultaneously, even across a network of human and algorithmic actors [44, p.11-12].

# 3 Legal principles, concepts and considerations

## 3.1 The development of an ethical and legal foundation for AI

In any society in any country there are ethical values and rules relevant to or governing that society or country. In a global context there is a variety of legal instruments such as conventions, treaties and agreements that sets the basic legal framework for countries on specific topics, for example the Universal Declaration of Human Rights [47] which was adopted shortly after World War II to establish a common international standard for the protection of fundamental human rights.

The technological revolution during the last few decades changed the world and continues to drive development. The digitisation of a variety of services impacts the way in which we do business, trade and even participate in sport. Various international organisations have raised questions about the ethical underpinning of the use of artificial intelligence and the legal principles that should regulate it or at least provide some guidance for detailed regulation to be developed. A golden thread that appears in these statements, proposals or directives is that the development and use of artificial intelligence should be human centred.

### Assessment Frameworks

There is certainly no shortage of initiatives, manifestos and guidelines that aims to provide ethical frameworks intended to provide assessment guidelines or inform the ethical culture of any entities involved with the design and implementation of AI algorithms. To date, more than 160 different AI ethics guidelines have been published (for a constantly updated inventory of Ethical Guidelines found globally, see Algorithm Watch [1]. Most of these guidelines discuss an overlapping set of values, such as privacy, fairness or non- discrimination, transparency, safety, and accountability [31].

However, not many of these guidelines are practical in terms of how assessment should happen and ethical values should be built into the design, deployment and use of these AI systems. It is from this perspective that we chose two influential assessment frameworks for discussion and use. For their breadth of application and their usefulness as assessment frameworks, we use as primary sources the EU's Ethics Guidelines for Trustworthy AI [15] and From Principles to Practice – An Interdisciplinary Framework to Operationalise AI Ethics [31].

The European Commission's High Level Expert Group on AI (AIHLEG) published a framework document called the Ethics Guidelines for Trustworthy AI in 2019 [15]. They argue that trustworthy AI has essentially three components, namely:

- It must be lawful.
- It should be ethical.
- It should be robust, both from a technical and a social perspective.

This Trustworthy AI framework identified seven key requirements for all AI systems, namely:

- Human agency and oversight.
- Technical robustness and safety.
- Privacy and data governance.
- Transparency.
- Diversity, non-discrimination and fairness.
- Environmental and societal well-being.
- Accountability.

This ethics guidelines document is not the only international document of this kind that aims to provide guidance to the development of ethical and trustworthy AI. Other important and useful international guidelines are the OECD Recommendation of the Council on Artificial Intelligence [49], the G7 Common Vision for the Future of AI adopted in 2018 in Canada [15] and the World Economic Forum's Framework for Developing a National AI Strategy [26], to name but a few. At the 40th International Conference of the Data Protection and Privacy Commissioners (2018) a declaration on ethics and data protection in artificial intelligence, that benefited from the vast practical experience of the contributors thereto, was adopted [33]. The Model Artificial Intelligence Governance Framework issued by the Singapore Personal Data Protection Commission is based on two guiding principles, namely (i) the use of AI in decision making should be explainable, transparent and fair, and (ii) AI solutions should be human-centric [56].

In addition, the volume of research about ethics and AI has increased significantly during the last few years, as a recent article from a research group at the Oxford Digital Ethics Lab clearly indicated [58]. The fact that various high level international meetings and organisations have debated the ethical and legal questions relating to the use of AI confirms the importance of a sound ethical and legal basis, and that this is a matter that requires international cooperation to develop a common understanding and approach to the ethical and legal foundations for AI. All these documents consider the design and use of AI in general and are not limited to a specific sector. In the context of this research study it is not only the general guidelines that are relevant, but it is also important to consider the ethical and legal issues pertaining to the use of AI in the health sector.

In the field of health care there are particular concerns in view of the use of personal information as part of the provision of health services. Patients need to provide their personal information to a doctor or other health care provider and can justifiably demand that such personal information is protected and only used in connection with the provision of health care to that patient.

During the Covid-19 pandemic many countries have introduced strict rules to fight the pandemic and those rules limited some individual rights with the aim of saving lives. An assessment of the design and application of these emergency or special rules justifies a study on its own. In this research study the focus is on the design and use of artificial intelligence in fighting a health pandemic such as Covid-19.

## 3.2 Key legal principles and concepts

The plethora of guideline documents regarding ethical AI lists a variety of ethical and legal principles, but there are some key principles that are found in most of these documents and they are discussed below. Most of the legal principles and concepts discussed in this section have their origin in constitutional law. This provides a useful point of reference to explore the meaning and scope thereof in the context of AI, in particular when using it in managing a health pandemic. Various international legal instruments such as the European Convention on Human Rights and the EU Treaties provide a sound legal basis for the further development of the law in the context of artificial intelligence. While this paper reflects on the meaning and scope of these concepts in the context of AI, each concept should be further contextualised to the specific field of application or use case.

### 3.2.1 Transparency

Transparency means to do something in an open and transparent way. Transparency is an important element of good governance and in a constitutional democracy necessary to strengthen accountability since citizens want to see and understand the reasons for government decisions and plans in order to hold them accountable. Transparency in the context of algorithmic decision-making is not that easy to determine. Ananny and Crawford, in their analysis of algorithmic accountability, came to the conclusion that transparency in this context provides limited help to explain and understand algorithmic decision-making [2] [11].

Transparency of algorithms includes two elements, namely accessibility and comprehensibility of information [44], which relates to the inscrutability of evidence. Lack of algorithmic transparency is influenced by various factors, such as the inability of humans to interpret large data sets and complex algorithms and the inherent opaqueness of machine learning algorithms. While transparency is an important ethical consideration that contributes to algorithmic accountability, some scholars argue that it should not be overemphasized [3] since it could potentially limit the focus of strengthening safety, performance and accuracy of the algorithm, and it could create an information overload.

A particular problem regarding transparency of algorithms is the time factor. Is it about a snapshot picture of the source code, an ex post facto view of the algorithm or the data sets? The problem increases in the case of machine learning algorithms that can adapt and change over time. A more realistic approach regarding transparency is needed to overcome some of these difficulties, namely 'looking into the system should rather be replaced by looking across the system in order to get a holistic view' [11]; [2].

In considering transparency as an important ethical factor, questions should be asked about the underlying concerns at different stages of the development of AI, for example how are the input datasets defined and obtained, when was the data collected, and what design features are used for the AI model? Explainability is perhaps a more accurate term to apply in the context of algorithmic accountability than transparency. Explainability is also contextual, namely different users might need different explanations [22]. Tsamados et al. [58] correctly argued that "Explainability is particularly important when considering the rapidly growing number of open source and easy-to-use models and datasets" .

| Principle | Description |
|---|---|
| **Fairness** | "Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics" |
| **Explainability** | "Ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms." |
| **Auditability** | "Enable interested third parties to probe, understand, and review the behaviour of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of detailed documentation, technically suitable APIs, and permissive terms of use." |
| **Responsibility** | "Make available externally visible avenues of redress for adverse individual or societal effects of an algorithmic decision system and designate an internal role for the person who is responsible for the timely remedy of such issues." |
| **Accuracy** | "Identify, log, and articulate sources of error and uncertainty throughout the algorithm and its data sources so that expected and worst-case implications can be understood and inform mitigation procedures." |

Table 1: FATML Principles (2016)

### 3.2.2 Accountability

Accountability can be a foundational principle of a constitutional system, for example in South Africa, and it form part of the system of checks and balances in a constitutional democracy [57]. The executive branch of government is accountable to the legislature which provides an important oversight role over the exercise of executive power. An independent judiciary fulfils an important function in a constitutional democracy by checking the exercise of power of the executive and legislative branches of government [18]. Accountability in this context means that the executive must take responsibility for its decisions and actions and report to Parliament (legislature). It is about explaining how the executive power was exercised.

When accountability is applied in the context of AI, it is not so straightforward and many questions such as who should be accountable for an algorithmic decision are raised. Accountability relates to transparency, as discussed above. Algorithmic accountability is about "the design and implementation of algorithmic systems in publicly accountable ways to mitigate harm or negative impacts on consumers and society" [27]. Busch pointed out that algorithmic accountability is not only about ensuring transparency, but that it also relates to the ethics of algorithms, the legal and technical requirements in their design, and societal considerations [12]. While the concept of someone being held accountable to consumers or society sounds simple, it is evident that algorithmic accountability is a much more complex concept that deals with various factors and limitations, for example the problems relating to transparency of algorithms discussed above. The principles developed by a group called the Fairness, Accountability and Transparency in Machine Learning community (FATML) are useful aids in constructing algorithmic accountability, namely fairness, explainability, auditability, responsibility, and accuracy. They describe these principles in Table 1.

It is evident from the literature that algorithmic accountability is not about an ex post facto once-off report, but it has rather a systemic application. The World Wide Web Foundation describes the comprehensive approach to algorithmic accountability aptly as follows: “Making algorithms more accountable means ensuring that harms can be assessed, controlled and redressed. Ensuring algorithmic justice implies finding the right remedies and identifying the responsible parties to take action” [27]. Such a systemic approach is supported by the Alan Turing Institute who argues that accountability should be applied over the whole algorithm life cycle and that answerability and auditability are the core elements thereof [38].

### 3.2.3 Dignity, equality and fairness

The right to human dignity is a basic human right recognised and protected in various international and domestic legal instruments, and it relates to the right to equality. De Vos aptly stated that human dignity is based on the notion that all humans have an equal moral worth, which means that legal protection of the right to equality is a logical consequence thereof [18]. The South African Constitutional Court has confirmed the importance of human dignity in relation to equality in a number of cases, and indicated that infringement of human dignity as well as other forms of differentiation that affect persons negatively and cause harm, could be in conflict with the Constitution [67]. Inherent in the right to human dignity is respect for each other as human beings, and this must apply to the use of algorithms as well. The AIHLEG argues that human dignity in this context requires that the intrinsic worth of a human being “should never be diminished, compromised or repressed by others – nor by new technologies like AI systems” [50]. It thus requires a human-centred approach in the design and application of AI, including how the input data sets are defined and collected in order to give effect to dignity. Although such an approach is commendable and legally sound, the possibility of conflicting views of the meaning of fairness in machine learning and the legal principle of fairness is recognised [65].

In the context of AI the issue of human dignity relates to the principle of fairness, which is aimed at ensuring that algorithmic decisions do not create discriminatory or unjust impacts or cause harm [24]. The AIHLEG links the principles of equality and fairness and stated that equality means the use of AI should not produce “unfairly biased outputs” [50]. This means that the input data should also respect the right to dignity and equality and the principle of fairness. This substantive dimension of fairness aims to ensure that individuals and groups are “free from unfair bias, discrimination and stigmatisation” [50]. Fairness must be considered in the particular context in which the AI is applied. What is fair and acceptable in one context could be unfair and unacceptable in another context. There is also a procedural dimension, namely that developers of AI should be able to balance competing interests and objectives in the design and application of AI. This would for example be the case in contact tracing apps to detect Covid-19 infections, where there are competing interests of privacy, safety and health care.

The Alan Turing Institute dissected the fairness principle as follows:

- Data fairness — representative, relevant, accurate, generalisable datasets.
- Design fairness — models should not be unreasonable, unjustifiable or morally objectionable.
- Outcomes fairness — no discriminatory or inequitable impacts on the lives of people the AI affects.

| Description | Legal provision | Legal document |
|---|---|---|
| Everyone has the right to respect for his private and family life, his home and his correspondence. | Sec. 8(1) | European Convention on Human Rights, 1953 |
| Everyone has the right to respect for his or her private and family life, home and communications. | Sec. 7 | Charter of Fundamental Rights of the European Union, 2012 |
| Privacy of correspondence, posts and telecommunications is inviolable. Privacy of the home is inviolable. | Sec. 10(2)<br>Sec. 13 (1) | Basic Law of the Federal Republic of Germany, 1949 |
| Everyone has the right to privacy, which includes the right not to have-<br>1. Their person or home searched,<br>2. Their property searched,<br>3. Their possessions seized, or<br>4. The privacy of their communications infringed. | Sec. 14 | Constitution of the Republic of South Africa, 1996 |

Table 2: Examples of privacy as a basic human right.

- Implementation fairness — users must be trained to implement AI responsibly without bias [38].

### 3.2.4 Privacy

The right to privacy is a basic human right recognised in various international legal instruments and country's constitutions, all of which defines this right in their own way. Some examples are listed in Table 2.

In the interaction between AI and big data the issue of personal information or data plays an important role. The right to privacy includes the privacy of personal information or data, as is evident from the abovementioned legal provisions. Protection of the right to privacy is aimed at preventing harm to individuals. AI could be very beneficial to society, for example algorithmic decision-making that can enhance human decision-making through increased efficiency and ability to deal with large datasets. In dealing with personal data AI could also cause harm for example by infringing the right to privacy. It is therefore necessary that processing of data, including personal data, is regulated to give effect to the right to privacy while also providing for the lawful processing of personal data.

Protection of personal data is the focus of the General Data Protection Regulation (GDPR) of the EU, which aims to

1. provide rules to protect natural persons when processing their data and standards for the free movement of personal data; and

2. protects fundamental rights and freedoms of individuals particularly, their right to personal data protection.

The GDPR views the right to privacy in a serious light and giving effect thereto in the context of data protection it created 8 specific rights of data subjects (individual persons), e.g. the right to be informed and the right of access to personal data. The right to privacy in this context does not entail an absolute prohibition against the processing of personal data but implies a balanced approach to protect privacy but also demarcating the space for lawful processing of personal data.

Data protection regulations are ordinarily designed to have a general application to the processing of data and are not specifically aimed at AI processes. There is, however, a specific provision in the GDPR to mitigate the risk of algorithmic decisions relating to personal data, which reads as follows:

> Art. 22(1): "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."

Mitrou argues that in order to understand the legal issues pertaining to privacy and data protection in the context of AI, there should also be a reflection of the public perceptions of AI and how AI are already used in our daily lives [43]. AI applications such as those used in speech recognition programs enhance human decision-making, but profiling through AI could potentially infringe on the right to privacy by for example interception of personal telecommunications. Various aspects of the right to privacy could be impacted by AI, for example informational privacy which means the ability of a person to control his or her personal information. In giving effect to the right to privacy it is therefore important to determine the degree of individual control over personal information in relation to the design and application of AI. This implies clear formulation of an appropriate legal framework based on the protection of the right to privacy, which will enhance trust of consumers to use the relevant AI applications. The Council of Europe emphasizes the importance of meaningful human control over the data processing by use of AI [16].

The right to privacy strengthens the protection of individual freedom and personal identity. This right applies to the processing of personal data in general and should also be given effect to when the data processing involves the use of algorithms or artificial intelligence. The responsible interaction with AI systems requires a human centred approach which includes the proper recognition of the right to privacy.

### 3.2.5 Other human rights considerations

In the previous sections the most important and commonly used legal principles and concepts that are important in ethical artificial intelligence were discussed. It is possible that other human rights considerations might, in particular situations, also be relevant. The AIHLEG suggests that the ethical principle of human autonomy, which relates to the right to individual freedom, should also be recognised in the interaction between humans and technology, including AI. The justification of the principle of human autonomy is described as follows:

"Humans interacting with AI systems must keep full and effective self-determination over themselves and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or hurt humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills" [50].

Recognising the principle of human autonomy implies a human-centred approach and some form of human control or oversight in the design and implementation of AI systems. The right to dignity, in addition to human autonomy, are affected if people are deprived from the right to exercise influence over decision-making processes that significantly affect them, for example in case of algorithmic decisions [43]. A legal provision such as Art. 22 of the GDPR provides a suitable response to give effect to the principle of human autonomy. Art. 8 of Convention 108 of the Council of Europe also determines that:

> "Every individual shall have a right: (a) not to be subject to a decision significantly affecting him or her based solely on an automated processing of data without having his or her views taken into consideration." [17].

Such an approach determines human involvement in the data collection and algorithmic processes. In the case of deterministic algorithms there is a clear human involvement that would recognise human autonomy, but with probabilistic algorithms used in machine learning it is more difficult to give effect to the principle of human autonomy. Machine learning is about the ability to learn without being explicitly programmed and this enables predictive modelling and complex problem solving being done by algorithms [43]. Machine learning would thus require a different approach regarding human autonomy compared to linear programming related to deterministic algorithms. This is discussed further below under "Process based legal considerations".

### 3.2.6 Health context

Artificial intelligence is regularly applied in health care, for example health apps for mobile phones that provide advice about specific medical conditions. For the AI models to work lots of personal data are required. This raises questions about individual rights such as privacy and dignity. Gathering of large amounts of personal health data, which is used to develop the algorithms, requires some form of consent of the patients in view of the right to privacy, including the protection of personal data. According to Nicholson the area of 'black-box medicine' in fact raises additional legal concerns like liability for damages and intellectual property regulation, but this is beyond the scope of the current research study [53].

During the outbreak of the Covid-19 pandemic in early 2020 various initiatives that utilised artificial intelligence in different ways were adopted, for example predicting the rise in infections in a specific area and contact tracing apps. While there could be significant benefits in utilising AI in fighting a health pandemic for example by reducing the burden on medical practitioners and hospitals [42], there are also various ethical and legal issues that warrant attention, for example privacy and protection of personal data. The GDPR includes some exceptions against the general prohibition of processing of personal data (Art 9(2) GDPR), for example if processing is necessary for the development of preventive or occupational medicine, or if processing is in the public interest in the area of public health, which clearly applies to health pandemics. The concept of public health should be interpreted widely to include for example issues related to health care services, financing of health care, disease control and managing pandemics. The processing of personal data should still be limited to the public health care issues and may not be processed outside the scope of this provision. This means that in case of public health issues there are competing ethical and legal principles that require a balanced approach to enable the lawful processing of personal data as well as to design AI models to respond to the particular public health issue.

The use of AI in managing the Covid 19-pandemic has raised human rights and other legal concerns such as the need for accountability and the right to privacy. While considering the various aspects related to the use of algorithms in the health context, it is apt to also reflect on what constitutes health data and how it could be legally processed. Health data includes:

- Information of a person collected by a health care provider,
- Information that becomes health data by cross-referencing with other data,
- Information from a self-check questionnaire about a person's own health, and
- Information that becomes health data due to its usage in a specific context, e.g. regarding Covid-19 contact tracing [62].

In processing such health data, the requirements of Art. 9 of the GDPR or of other applicable regulations depending on the particular jurisdiction will have to be adhered to. This means that the designers of an algorithm that would be applied to the input data have to ensure that the legal requirements for the processing of the data are met. In a recent publication by the Council of Europe the importance of adherence to the relevant data protection regime, e.g. GDPR, when processing personal data in fighting a health pandemic was emphasized, namely:

1. The need for a time limit for the retention of the data.
2. Clarity and a sound legal basis for the purpose for which the data is processed.
3. Proportionality of the measures taken in processing the personal data.
4. Transparency and explainability of the data processing.
5. Accountability.
6. Integration of privacy by design.
7. Realisation of data protection impact assessment [59].

Data protection regulations are thus one of the key determinants that will guide the ethical use of AI in managing health pandemics.

### 3.2.7 Governance

A discussion on ethical and legal issues relating to artificial intelligence would not be complete without exploring the governance of AI. Governance can be defined widely as "the system of values, policies, and institutions by which a society manages its economic, political and societal affairs" [55]. In the context of artificial intelligence and this specific research study, it is argued that an adapted definition of governance that fits the AI context would be more appropriate, for example: AI governance consists of the ethical values, legal principles and oversight mechanisms for responsible or trustworthy AI. The Singapore Model AI Governance Framework suggests that internal governance structures in an organisation could contribute to effective oversight over the use of AI [56]. In terms of the Framework for Trustworthy AI proposed by the AIHLEG governance is a crucial element throughout the lifecycle of an AI system, for example respecting human rights and ensuring effective human oversight and adhering to privacy and data protection provisions.

Governance is thus not about compliance to a set of rules, but rather about acknowledging and giving effect to the critical interrelated elements of trustworthy AI applied in a specific context, e.g. in managing health pandemics. The element of human oversight as part of the governance architecture warrants some explanation. While it is not always possible nor feasible for human intervention in the use of AI, human oversight should be considered at various stages, namely in the design phase of the AI, the monitoring of the AI system and the overall application and impact of the AI system. This is discussed further in the next section.

## 3.3 Process based legal considerations

The Bayerisches Forschungsinstitut für Digitale Transformation (bidt) proposes an agile approach in dealing with ethical considerations in the development of algorithms [9], which implies that ethical considerations should be considered throughout the process of development and use of algorithms. This would also apply to the legal considerations in view of the interwoven nature of ethics and the law in this context, although in case of the law the developers of AI would not only consider the legal implications, but also have to ensure compliance with relevant regulations. This is in line with the systemic approach to the development of technology described by the World Economic Forum, which suggests that values and ethics of society should be considered in all the phases of development of technology [52]. There is in fact an interrelationship between ethics and law which contribute to the shaping of the architecture of artificial intelligence. Some scholars argue that an ethical analysis of the use of AI in a particular context is necessary to identify and mitigate the risks and to harness the potential for doing good, something which is also relevant in managing health pandemics [58].

When there is clarity on the ethical and legal foundations for an AI project, the question is how should that be incorporated into the whole AI life cycle? How can the principles be operationalised?

“Design for Values” is a methodological process which includes the identification of societal values, deciding on a moral deliberation approach (e.g. through algorithms, user control or regulation and linking values to formal system requirements and concrete functionalities [20].

The Design for Values approach includes:

1. Identify the relevant stakeholders,

2. Elicit values and requirements of all stakeholders,

3. Provide means to aggregate the values and value interpretations from all stakeholders,

4. Maintain explicit formal links between values, norms and system functionalities that enable adaptation of the system to evolving perceptions and justification of implementation decisions in terms of their underlying values, and

5. Provide support to choose system components based on their underlying societal and ethical conceptions, in particular when these components are built or maintained by different organizations, holding potentially different values [20].

What Design for Values mean for the design, development and deployment of AI systems, is that the usual development cycle (analysis, design, implementation, evaluation and maintenance) cannot be adequate. Evaluation and ethical justification should be part of every step of the process [20] .

The assumption is that values are embodied in technology as the result of certain intentional value-embedding activities by designers. The challenge with AI systems is that it carries the building blocks of socio-technical systems (namely technological artefacts, human agents and institutional rules), but that it has the additional capacity to autonomously interact with its environment and adapt itself on the basis of such interactions and learning. This may result in, perhaps unintended, disembodying of values that were originally embedded by the system designers [61, p. 387]. Therefore, it needs to be realized that besides human agency, artificial agency also plays a role, as does the technical norms to be embedded in the system. Embodying values means that the system has to be intentionally designed to comply with that value and that the system has to further that value when it is used properly. "The embodied value is the value that is both intended (by the designers) and realized if the artefact or system is properly used" [61, p,389].

Establishing rules is one way to intervene in an AI system to embody the values via the relevant artefacts. Also, human actions are required for the proper functioning of sociotechnical and AI systems. Human agents can be seen as being a part of such systems, fulfilling various roles such as user, operator, and designer. It is precisely in the interplay between human and artificial agents, that human agents not only embed values in technological artefacts, but also evaluate the outcome to confirm that the values are realized and adapting the rules and code in order to prevent disembodying of the values. An artificial agent disembodies a value if it adapts itself in such a way that it is no longer conducive to the value, even if it originally embodied that value [61, p. 400].

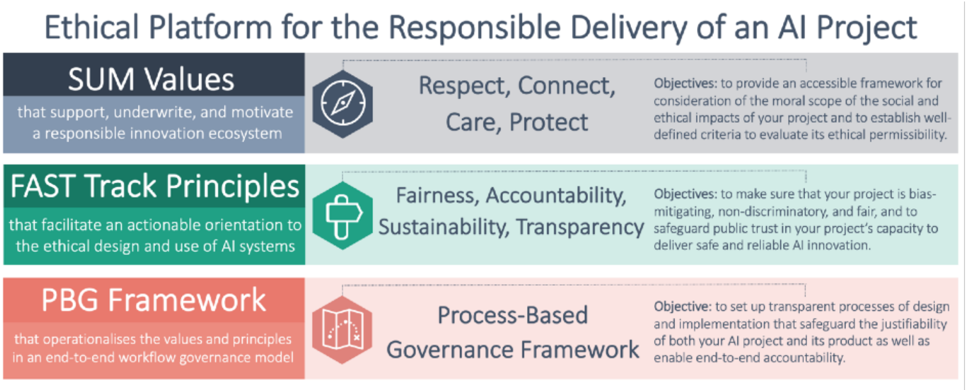
The implementation of ethical reasoning can be divided into three main types [20, pp. 75-81]:

1. **Top-Down Approaches.** Here individual decisions by an AI system are inferred from general rules. A given ethical theory are prescribed to the computational system and applied to the particular case. The difficulty lies in deciding which ethical value should be maximized for. Is it fairness, human dignity, trustworthiness? Because maximizing for different values may lead to different results. A utilitarian view will maximize "the best for most", while a deontological view may want to evaluate the goodness of actions. And what, in any case, is the "best" – is it wealth, health, sustainability or some combination of values? It also cannot mean that legal codes are merely enforced on AI systems, as ethical systems need to decide what ought and ought not to be done over and above existing regulation.

2. **Bottom-Up Approaches.** Here general rules are inferred from individual cases. Given sufficient observations of what others have done in similar situations, the system can derive what is ethically acceptable. The system builds up, through experience of what is to be considered ethical and unethical in certain situations, an implicit notion of ethical behavior [13, p. 179]. The problem is that societal acceptance of behavior is equated with ethical behavior here. The "wisdom of the crowd" can potentially lead to accepted but unacceptable decisions [20, p. 80].

3. **Hybrid Approaches.** This combines elements from top-down and bottom-up approaches to support careful moral reflection which is essential for ethical decision- making. Rules as well as context observations are needed and used to implement ethical agents.

The Turing Institute proposed a process-based governance model that span the AI life cycle, which includes an ethical foundation and incorporation of key legal principles [37]. The process flow diagram in this proposal is depicted as follows:


Ethical Platform for the Responsible Delivery of an AI Project
SUM Values
that support, underwrite, and motivate a responsible innovation ecosystem
Respect, Connect, Care, Protect
FAST Track Principles
that facilitate an actionable orientation to the ethical design and use of AI systems
Fairness, Accountability, Sustainability, Transparency
PBG Framework
that operationalises the values and principles in an end-to-end workflow governance model
Process-Based Governance Framework

**Objectives:** to provide an accessible framework for consideration of the moral scope of the social and ethical impacts of your project and to establish well-defined criteria to evaluate its ethical permissibility.

**Objectives:** to make sure that your project is bias-mitigating, non-discriminatory, and fair, and to safeguard public trust in your project's capacity to deliver safe and reliable AI innovation.

**Objective:** to set up transparent processes of design and implementation that safeguard the justifiability of both your AI project and its product as well as enable end-to-end accountability.

Figure 1: Ethical Platform

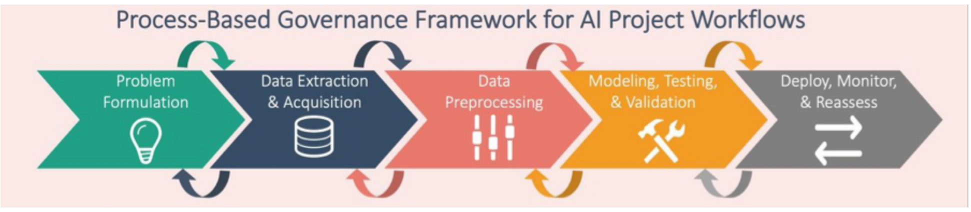

Process-Based Governance Framework for AI Project Workflows
Problem Formulation
Data Extraction & Acquisition
Data Preprocessing
Modeling, Testing, & Validation
Deploy, Monitor, & Reassess

Figure 2: Process Flow

### 3.3.1 Acquisition of data

Very often data sets reflect an inherent bias, and if this is not reviewed it will have an impact on the design of the algorithm and its output when it is applied to the input data sets. The principle of fairness requires that such unfair bias should be avoided. The following key elements of data fairness should be considered:

- Representativity
- Sufficiency
- Integrity and accuracy
- Timeliness
- Relevance, appropriateness and domain knowledge (Leslie, 2019).

The objective in defining the dataset to be used as input data is to support the principle of discriminatory non-harm. This means that the input data must be properly representative, accurate, relevant and generalisable (data fairness). When this is applied in the context of this research study, namely managing health pandemics, it means that the simulated datasets, i.e. for diseases, long term social networks and mobility, should adopt such a data fairness approach.

When the simulated datasets are tested against real data, care must be taken to mitigate bias and ensure data fairness as well. Real data must further be relevant, accurate, appropriate and up to date. Privacy remains a key concern and is part of the data protection legal architecture.

It is further important to note that transparency, accountability and fairness are not only legal principles that guide the development of the AI model, but they are also legal requirements that apply to data processing as stipulated in the GDPR or other legal frameworks (see previous sections regarding data protection requirements in health care). Following the Turing flow diagram, it means that the data processing will be measured against the data protection requirements at various stages of the AI process.

### 3.3.2 Design of AI

Fairness in the design of AI models relate to the following elements:

- Problem formulation
- Data pre-processing
- Feature determination and model building
- Evaluating analytical structures [38].

Competing legal principles will often occur due to the variety of human needs and interests that exist in society. However, it is perhaps during this part of the AI life cycle where it is most important to have a pragmatic approach that allows for a balancing of competing legal principles. This is for example the case with predictive policing where there are competing rights of human autonomy, privacy, freedom of movement and the principle of prevention of harm. A reasoned and evidence-based approach is necessary in finding an acceptable balance, which would also depend on the particular context [50].

In the context of managing a health pandemic, governments have to take difficult decisions that could infringe some human rights such as freedom of movement with the aim of saving lives. This means that there is a weighing up of individual rights. When AI is used as a tool that assists in the decision making, the balancing of competing interests and different individual rights extends to the design of the AI and how it is applied in practice.

The initial problem formulation should be translated into specific measurable targets that will influence the design of the AI model, for example:

1. An inclusive checklist of legal criteria that will promote responsible and trustworthy AI, and
2. Indicators for spreading of an infectious disease within a particular community or geographical area based on critical factors such as social interaction, mobility of people and existing extent of infectious diseases.

The labelling and classification of the input data must adhere to the fairness principle, with due consideration to the relevant societal contexts [37]. Ethics by design is a specific approach that will contribute to achieving responsible AI.

In view of the importance of the principle of accountability, the Turing Institute suggests an accountability by design approach, namely that there is a conscious decision to incorporate accountability in the design and throughout the life cycle of the AI, that will facilitate answerability and auditability which both contribute to accountability [38].

The legal principles and other key requirements for responsible or trustworthy AI should be translated into the technical process and architecture of the AI model. In case of deterministic algorithms, the monitoring of compliance with these design features can be done in the next phase of the AI life cycle by way of a dedicated monitoring process. In case of machine learning AI models, the monitoring is more complex and should be dealt with in a different way.

Designing AI models to assist in managing health pandemics is a multi-disciplinary task and policy expertise and domain knowledge is an important contribution to the technical development work. It is at this stage that careful consideration should be given to mitigate potential risks and unfair bias that could occur. The design process should be an iterative process that includes testing against the set legal requirements and other design criteria. Such an approach will contribute to ensuring accountability. In terms of the process-based governance framework proposed by the Turing Institute, design fairness also includes assessment to detect hidden proxies for discriminating features in the structure of the AI model.

### 3.3.3 Application of AI

Procedural fairness should apply throughout the AI lifecycle. This means that ethical considerations, procedures and legal requirements should be applied in a consistent manner. In the context of health pandemics, it would mean that the AI model must be designed in such a way that it can be replicated in a procedurally fair way.

One of the key legal principles in support of responsible AI is transparency or explainability. With this in mind, the description and parameters of the input data should be clearly described to provide a first level of explainability about the process. The transparency requirement is more difficult when dealing with machine learning due to the nature of the algorithmic processes. It should first be ascertained what kind of explanation for the AI model is needed. A description of the logic behind the design of the AI model, its goals and how it can achieve these goals might perhaps satisfy the explainability requirement in this health care context. The outcomes of the application of the AI model should be explained as part of giving effect to the principle of transparency. Explainability is not only about the technical process but also about the related human decision-making process.

An important part of the development process is testing and validation of the AI model to test its stability and robustness. This is a technical process, but it must still meet the legal requirements for responsible AI.

Accountability is one of the key legal principles that underpin responsible AI. It includes auditability, meaning that an objective assessment of the data, design process and the algorithms should be possible. It is during this phase of the AI life cycle that conflict between different interests, ethical or other requirements are clearly identified. It is inevitable that some trade-offs need to be made to refine the AI model, but this needs to be done in a methodical and reasoned way in order to satisfy the accountability requirement [50]. Part of assessing accountability is also to ensure that redress is possible if something went wrong or unfair outcomes are produced. This could strengthen the trust in and legitimacy of a specific AI model.

### 3.3.4 Impact of the use of an AI model

In managing health pandemics different kinds of AI applications could potentially be utilised, for example contact tracing apps and robotic medical assistants. This current research project focuses on other possible use cases, namely where predictive modelling is used to assist decision making in managing health pandemics. An important phase in the AI life cycle is to review its impact within the particular context it is used. Outcome fairness is based on the human rights concept of fairness, namely that it should not cause harm, discriminatory or inequitable impacts. The application of AI in different contexts, e.g. in self-driving cars, search engines or health care, relates to a varying degree of risk. While the basic point of departure in ethical AI is to follow a human-centred approach that will lead to fair outcomes, it is necessary to assess the potential risk of an AI model. In applying the critical ethical and legal principles to the AI life cycle, an important question is how is it done in a logical way that could be measured objectively? Simply put, how are the principles translated into practice? The AI Ethical Impact Group (AIEI) developed a comprehensive practical model which applies six key principles, namely transparency, accountability, privacy, justice, reliability and environmental sustainability, to the development and implementation of AI [29].

It provides a very practical approach to consciously focus on how these principles are applied and indeed provide some measurement of the application as well. Once the principles are applied and the AI model is implemented, there could still be unanswered questions about the impact or outcomes of the particular model. It is therefore important to have some form of risk assessment. The AI Ethics Impact Group (AIEI Group) provides a useful tool that focuses on potential harm and degree of exposure of affected persons to the AI model. They propose a risk matrix:

- Intensity of potential harm (x-axis), which considers impact on human rights such as equity, fairness; number of people affected; and the impact on society.
- Dependence on the AI decision (y-axis), which considers the degree of human control; switchability between different options; and the possibility of challenging or correcting an algorithmic decision (redress) [29].

# 4 Technical project design and development

## 4.1 Data simulation and research overview

To facilitate and enable better decision making with respect to ethical and legal principles in governmental response to a pandemic scenario, part of the project's focus was to provide a structured framework to test the applicability and efficacy of machine learning-based models for disease spread prediction. Since public health data for this type of scenario is not only highly sensitive and most often incomplete, it was important to deploy a framework that is able to generate simulated disease spread data under different parameters.

This simulation environment is at the core of the technical framework for the project and it fulfils two main purposes:

- Firstly, in the absence of widely available data for disease spread, the simulation can generate large amounts of data necessary for neural machine learning approaches reliant on such data.

- Secondly, the simulated data affords the possibility to range from very general statistics about the simulated spread, down to very fine-grained data about each infected individual and their movements in the system. This is also vital, since it allows to test on wide range of possible input features for Machine Learning methods, which in turn allows to efficiently select only the most relevant data for the prediction task.
- Thirdly, the simulation allows feedback from our prediction system, to test the effects of the models predictions on the system, to judge their effectiveness in reducing the spread of the disease.

Through this simulation framework it is then not only possible to obtain the data for the training of the model quickly and efficiently, but most importantly it enables us to test whether the collection of such data would be helpful for the actual task before setting out to collect it. This preliminary testing of a methods potential effectiveness can provide essential insights into whether the collection of such data for disease prevention can be justified ethically and provide legislative decision makers with more background for their policy decisions.

During the project, work on the technical part was done in several smaller project parts, focusing on different parts of the simulation pipeline and on the methodology with which to tackle the problem of disease-spread prediction based on the simulated data. The following sections offer an overview of each of these parts of the project with a more in-depth description of the task, assumptions and outcomes.

## 4.2 Disease spread simulation environment

In order to ethically obtain disease spread data that can be used for Machine Learning purposes, a modular, flexible and efficient simulation had to be developed. The simulation approach discussed in this report separates the general simulation cycle into three main modules: Long Term Social Network, Mobility/Encounters and Disease Spread.

In the Long Term Social Network, strong social ties between nodes as well as the assignments to their according work, education and home facilities are created. Generally, a single social tie falls under one of the following categories: friend, family, household, partner, co-worker. Additionally, a node profile is stored for every person in the simulation, containing information about that specific node's infection status and various personal attributes such as age or individual sociability. Also stored in the Long Term Social Network module are additional customizable facilities that nodes are able to attend during their free time, as well as user-defined means of transit.

The Mobility module creates a weekly schedule for every node. These schedules depend on whether a person is employed, attending school and the count of activities they partake in during their free time. These schedules consist of 48 30-minute slots per day. These schedules allow the simulation to determine in which facility a node is present at any given time. That information combined with the sociability scores of each node is used to create daily encounters for every person. An encounter is defined via its duration and intensity. They are split into three categories: connected, scheduled and random. Connected Encounters are the planned encounters a person has with their long term social ties. Examples are dates with one's significant other or a night at the cinema with one's friends. These are typically the most intense and longest types of encounters. Scheduled encounters are determined via the schedule of each node. If two

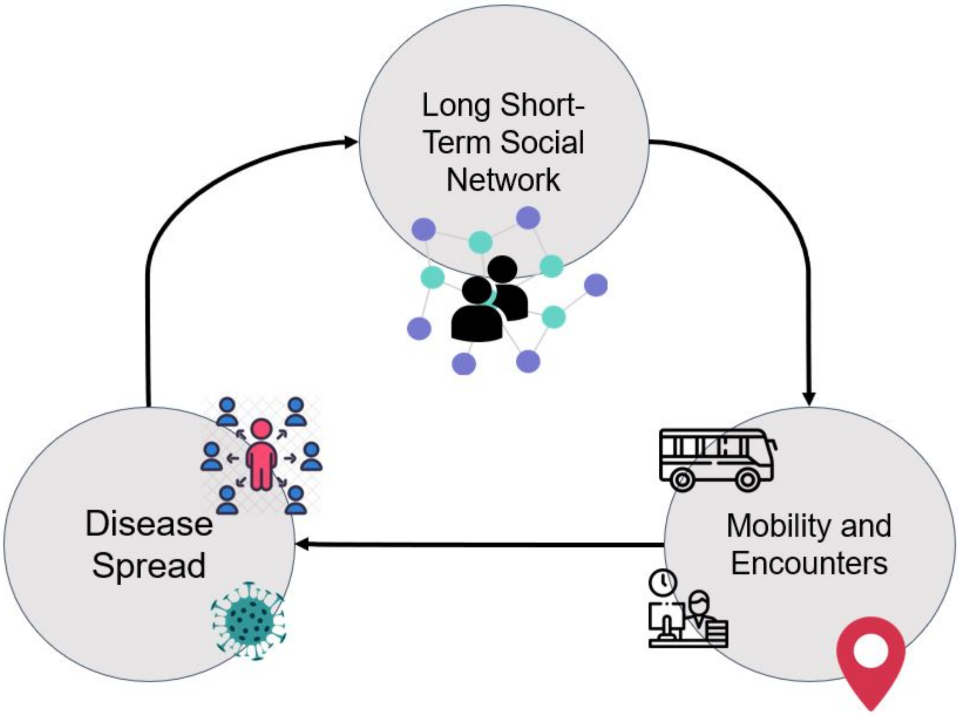


Figure 3: Three constituent parts of the simulation framework.

people are present in the same facility at the same time slot, they have a certain probability of interacting with each other, e.g. talking with a stranger on the bus ride to work. For these, the duration and intensity are typically lower than for connected ones. Random encounters account for the rigidity of the scheduling process and allow for interaction between nodes who would never be present at the same facility. An illustrative example is Person A eating at an outdoors table of a restaurant, while a randomly passing Person B asks them for directions to the main train station. In this case these two people would not be present in the same facility yet they are still able to interact. Random encounters usually have very low intensities and short durations. For each encounter, all the necessary meta data is stored, such as facility of the encounter, city district, time and the involved nodes as it will be important for learning purposes.

The disease spread module takes daily encounters as input and computes a probability for each infection of leading to transmission of the disease. Here only encounter duration and intensity are significant variables for the probability calculation as every other simulation parameter directly influences the former two. This module then decides for each encounter whether it led to transmission. While high risk encounters most likely will lead to infection, this allows for a more realistic uncertainty. That way, low risk infections are also able to contribute to the disease spread. After the long term social network creates and updates the node profiles it passes the node information to the mobility module. The latter then computes the schedules for one week in advance and creates encounters for the current simulation day. The encounters are then passed to the Disease Spread module, which determines which nodes are newly infected after the current simulation day. This updated infection data is given to the long term social network, which in turn updates the node profiles and the cycle repeats. Additionally the simulation allows for several restrictive measures to stop the spread of the disease, such as mask wearing, which reduces the average intensity for most activities, quarantine process, which sends nodes that show symptoms after being infected into quarantine. Quarantined nodes are assigned a special schedule by the mobility module to minimize contact to other nodes. Finally, after a specified number of simulation days, the generated output data, namely daily encounters and infection data of all nodes, can be used for further learning purposes.

For more details, see the full work from Drothler [21].

## 4.3 Simulation and mining of social networks to predict individual infection risks

The German Corona Warn app can notify its user if they had contact with a person who was infected with the Coronavirus. However, this notification only works if the infected person enters a positive test result in the app. We propose an extension to the warn app with a live update feature, that does not need a positive test result but can update the risk in place based on the encountered peoples' risks. With this, contact with a potentially infectious, i.e. high risk, person already increases the individual infection risk. In addition, we enhance the prediction scale from a binary prediction (high or low risk) to continuous infection risk.

For every day, we organize the people meeting on a respective day as an encounter network around the app users. Every node in this network is a person with individual infection risk and en edge indicates an encounter between the two people. An encounter has multiple features, such as the duration and facility type where the two people met, as well as the relationship the two people share. We propagate the infection risks through the encounter network using a page-rank-inspired approach. For this, we employ a three-step pipeline. First, we build the individual encounter networks for every person. We want to process the risk updates decentrally, i.e. only the phones of the people meeting know about the encounter. Therefore, we do not generate the full network, but individual excerpts of it for every user. Secondly, we train a regression model that takes the encounter features as input and predicts if and how infectious this encounter was. With this model, we can weigh each edge in the network by its infectiousness. To retrieve the individual infection risk for a single day, we sum over all encounters of that day. This sum is limited to be between zero and one to generate a proper probability. We store the daily risks of the last 14 days. In the end, the risk history is aggregated into the user's current infection risk using another regression model.

With this extension framework, we can not only improve the risk prediction of the warn app but also deduce implications for the real-life application of the app. For example, we used the first regression model, which predicts the infectiousness of an encounter, to investigate the important features of an encounter, e.g., whether it makes a difference if two people meet at home or work. Moreover, we try the prediction framework with different app covers among the population, i.e., we omit random users that do not use the app. We find that more than 75% of the population must use the app to yield reliable predictions.

For more details, see the full work from Anschütz [4].

## 4.4 Privacy-preserving differentially private models for disease prevention

Extraordinary effort is being taken to incorporate AI technology to fight the current COVID-19 pandemic. However, for this technology to be effective, massive amounts of sensitive data from the population are needed. In non-authoritarian countries, this requires the consent of its citizens for the use of their personal data.

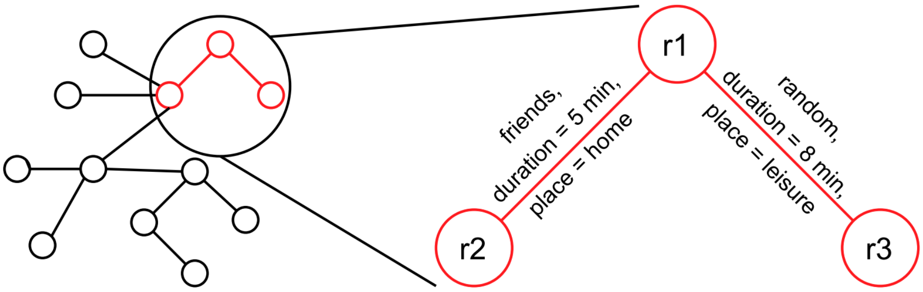


Figure 4: Encounter network of one day and a closeup of three nodes in this network

In this setting, this research aims to create an AI model that would help fight a pandemic and would have the trust of society by bringing AI and ethics together. In order to do that, this project creates a differentially private and explainable deep learning model for individual risk prediction and provides an evaluation of its level of trustworthiness using the Assessment List of Trustworthy AI and the Ethics Guidelines for AI produced by the European Commision's High Level Expert Group on AI (AIHLEG). Moreover, this project provides an analysis of the trade-off existent between privacy, explainability, and performance.

For creating an explainable deep learning model, the technique DeepSHAP is used. SHAP aims to explain individual predictions by measuring the contribution of each feature to the output using the cooperative game theory concept of the Shapley Value and adapting it to the game theory concept of machine learning. With DeepSHAP, we calculate the SHAP values offering a better computational performance in neural networks than any other model agnostic method.

For creating a differential private deep learning model, two papers [51] [68], are combined and adapted to our specific application, and the derived mechanism is proved to satisfy differential privacy at the same level.

The resulted mechanism: 1) adds noise to features depending on their relevance to their output, 2) can be used in all kinds of neural networks and 3) its privacy loss doesn't depend on the number of training steps. A differentially private loss function is also constructed using the Functional Mechanism [68].

Our results showed that by creating a differentially private model, its performance turns out to be precarious at any level of loss of privacy. Furthermore, the explanations provided by the model are also highly unstable and unreliable. Finally, the trustworthiness evaluation showed that while the differentially private and explainable model could satisfy the requirements of Transparency, Privacy and Data Governance, and Accountability from the Ethics Guidelines for AI, it failed to satisfy the other four ones mainly due to its poor performance and its high instability.

All of this concludes that it is not possible to build a trustworthy model for individual risk prediction. However, some limitations exist in the simulation created and in the mechanism for achieving differential privacy. The numerous benefits that offer to close the gap between AI and ethics can only stress the necessity and motivation for further research in this topic by addressing and solving the limitations of the project.

For more details, see the full work from Lopez [39].

### 4.5 Decentralized machine learning and federated learning for privacy-preserving predictions

We want to predict the infection risk using centralized learning methods, in which we have access to all the data, and decentralized learning methods, where data remains in distributed nodes that also help with the training process.

The models that we have used are the following:

- A random forest model with 300 decision trees as a baseline model using the default parameters of scikit learn framework.
- A centralized model called DPBoost which uses the ensemble base method called Gradient Boosting Decision Trees (GBDT) and differential privacy. Each tree is trained on the misclassified examples of the previous ones. Since we have to do with private data, we have also used this model that incorporates the differential privacy part which guarantees that the probability of producing an output does not change even if we include a record or not.
- A federated learning model as a decentralized model, where model parameters are trained on multiple devices and are then iteratively combined in a global model, in which we do not have access to the data, which serves as an advantage especially for sensitive data such as healthcare data. Each distributed node has a local copy of a multilayered perceptron with the latest trained parameters that are aggregated from the global model.

We have used grid-search for hyper-parameter tuning for both DPBoost and Federated Learning. We have used F1 Score as the metric as can be seen in the following table. Overall, these relatively poor performances of both models can be attributed primarily to the hyperparameters that were used which is generally a challenge on its own, especially for federated learning. Using additional features in the dataset could have arguably helped both models.

| **Model** | **F1 Score** |
|---|---|
| DPBoost | 0.6506 |
| **Federated learning** | **0.6673** |
| Random forest | 0.60 |

Table 3: F1 Score of machine learning models, assuming 66% of infected people are already known

For more details, see the full work from Morina [46].

### 4.6 Simulation-based management testing

To combat the pandemic spread, an individual risk prediction model and some Covid-19 preventative measures are tested in simulation. All the strategies are adapted into the simulation and therefore they has direct impact on the outcome of each simulation run. For instance, banning specific activities or locking down specific days have nearly no effect since they just slow down the spread but cannot stop it. In the following, our three most successful ideas are shortly described.

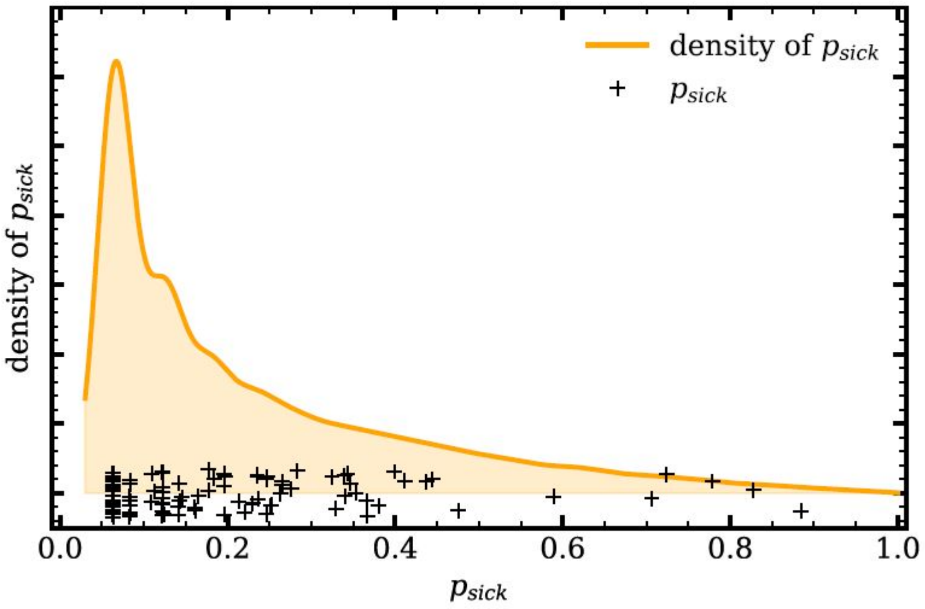


Figure 5: Risk of encounters leading to infection in the model

First, just use backtracking. The general idea of this method is that each person that has symptoms will be quarantined, and based on their encounters the people with direct or indirect contact will be tested and as the case may be quarantined. This works similar as the German corona warn app. If a person $p_i$ has contact to a infectious person $p_j$, the person will be notified. Even people without symptoms in most cases can be backtracked and the infection spread will stop.

Another way to stop the spread is to use the so-called ”2G” plan. Only when the person is tested or already recovered from the corona infection the person is allowed to take place in certain activities.

Third, with the usage of a simple recurrent neural network, the health state of each person can be predicted and then used to quarantine sick people. The general setup is the following: All encounters of one person are considered as a sequence. Each encounter is embedded in a vector and processed step-by-step by a recurrent network. For the embedding, the health state of the other person from the previous predicted day, the intensity and duration of the encounter is used. In the end, the last hidden cell is used as input in a feed-forward neural network to predict the new healthy state of this person. Unfortunately, this would lead to a high number of false positives. To reduce this number, each sick predicted person with a low infection risk will be tested.

Of course, the previous described methods only work in certain circumstances. The first and third method need a godlike view of each encounter and even if 10% of the people do not use a tracker it results into a slow linear increase of infections. For the ”2G” plan, the tracker is not needed but the most people must abide by the regulations and it should be used by many different activities. Of course, the combination of these methods is as well possible and might cause a slower spread.

For more details, see the full work from Fischer [25].

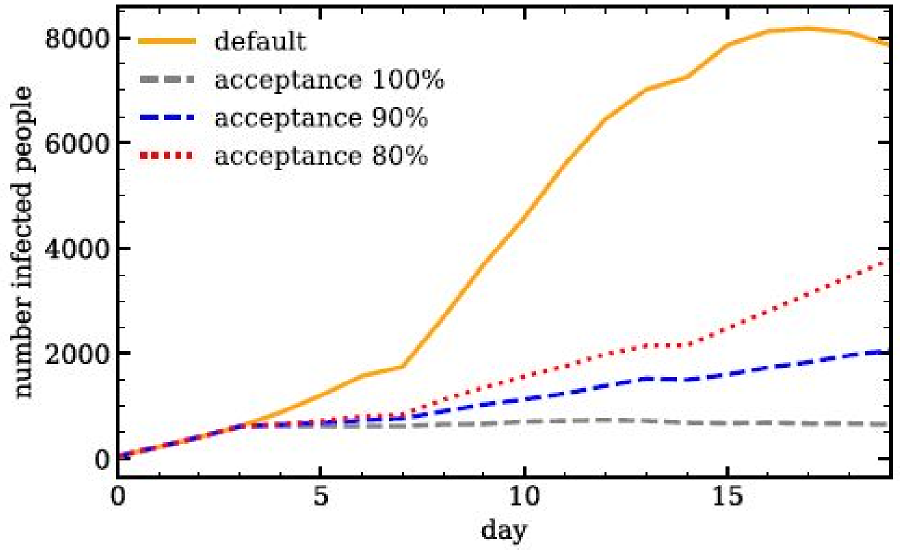


Figure 6: Development of infections in the simulation at different rates of compliance.

### 4.7 Risk prediction based on spatio-temporal data

Infections occur near one another. In this setting, we use data about facilities, as well as infection events, to find hotspots using clustering algorithms. Hotpots may not remain stable over time and the temporal nature must be considered. We combine these hotspot trends over time to get spatio-temporal hotspots.

Next we use hotspot information, such as the percentage of time periods classified as a hotspot and the trend to become a hotspot, to predict risks of citizens catching the virus without using their personal or private information. In this project we experimented with and will continue to research

- Predicting the individual risk of catching the virus after visiting a facility at specific point of time.
- Estimating the cumulative risk of being infected at the end of the day after visiting multiple facilities.
- Identifying and predicting hotspots and high-risk areas.

The resulting models can help citizens to avoid visiting high risk areas, estimate their risk of being infected and get tested if necessary. Furthermore, governments can use this information to selectively regulate the districts and facilities that are hotspots or are predicted to become one. For modelling we currently use the district information where the visit is happening, the facility type, how crowded usually at that time period the facility is, how many cases are there in the city and how long the person stays (or plans to stay) in the facility. The features are chosen to be informative enough to get a decent prediction, but also general enough not to contain identifiable information. At this point the models yield high true positive and true negative, and low false negative rates, which ensures that most of the positive cases can be predicted. However, the models also yield high false positive cases, which can result in a high amount of unnecessary testing. In the next stage of the research, one of the goals is to bring the false positive rate down, and hence, reduce the number of required testings.

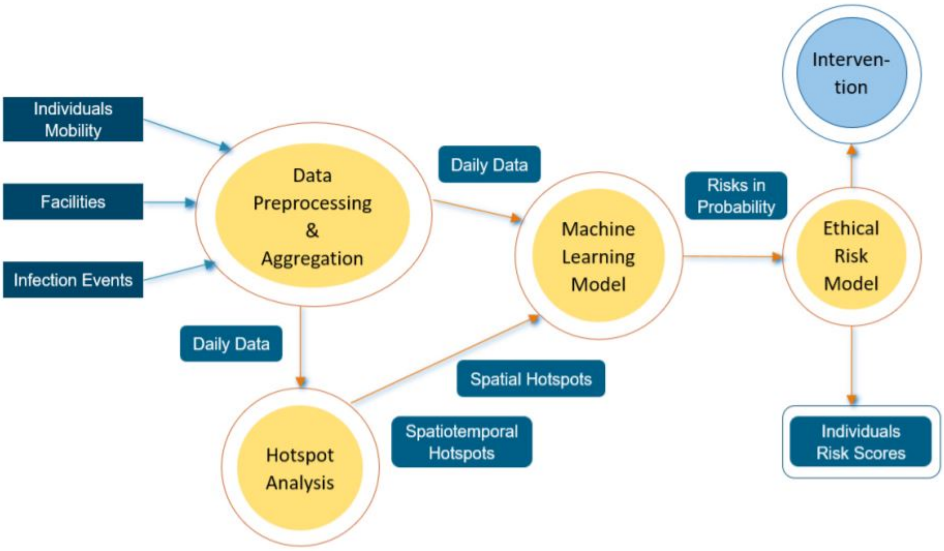


Figure 7: Pipeline steps for a spatio-temporal risk prediction model.

We also investigate intervention scenarios if users act on the hotspot information. In particular, we reroute 15% of visits from hotpots to facilities that are not hotspots. According to our simulation, this leads to fewer infections.

For more details, see the full work from Bezirganyan [8] and Lee [36].

### 4.8 Risk prediction based on personal activity and schedule evaluation

With the help of spatial-temporal information added to our agent-based simulation (including facility, district, ring, as well as the timeframe where human-to-human interactions occurred), activity-based prediction models can be built to analyze an individual's risk based on one's mobility behavior. An activity can be considered as a set of interactions taking place over some time at a particular location. In our simulation, the default activities include work, education, shop, cinema, gym, bus, train, car, restaurant, and home. To predict activity risk, the encounter data are first aggregated and transformed onto an activity level, where personal schedules can be observed and analyzed. Furthermore, the ground truth labels are mapped to represent the riskiness of an activity based on the number of infectious encounters. The influence factors for the risk of an activity instance can be grouped into three categories, namely, general activity risk, area-based risk, and other specific measures. General activity risk implies a static score, which proposes the relative infection probability for a broad activity group provided by the health officials. Area-based risk features describe how risky a region or district is, including, for instance, crowdedness, hotspot, and the effectiveness of social distancing. Lastly, other specific measures regarding a particular activity instance may include, for example, the intensity, duration, and popularity of the facility during the visit time. In the end, contextualized activity risks can be predicted using Machine Learning techniques based on the above-mentioned groups of features. The same actions can be associated with different risks if they take place at various locations or times. As a result, individuals can plan their future travels to make informed decisions and minimize risk. Moreover, policymakers can model the what-if scenarios statistically. The effects of the possible interventions can also be measured quantitatively, to mitigate the disease spread and support public health decision making.

For more details, see the full work from Wu [64].

## 4.9 Trustworthy Risk Prediction

Our goal is to build an algorithm that can assist policymakers in a pandemic situation by providing actionable insights on a given population's state regarding the spread of the infectious disease. For this purpose, our algorithm maintains an estimated state of the population over time. In this context, a population state at some day $t$ simply refers to the collection of health statuses of all individuals in the mentioned population at day $t$. We differentiate between three such statuses: *healthy*, *infected* and *cured*. The algorithm takes the state at some initial day $t$ and predicts the new estimated state at day $t + 1$. It achieves this by taking into account all encounters between pairs of individual people that happen on day $t$. An encounter simply is an event of two people A and B meeting. We parameterize these events with certain features, such as *duration*, *intensity*, person and A and B's (estimated) *health status*, *the relationship type* between the two people, etc. We point out, that the same encounter usually produces different risks for A and B. If A is infected and B is not, then the resulting risk of an encounter between the two will be higher for B than it would be for A. This is due to the underlying assumption that we have made in our simulation, that only infected people may potentially infect only healthy individuals. At the end of day $t$, the individual encounter risks are then aggregated for each person to produce their new risk at day $t + 1$.

Our algorithm could support decision making by helping to guide scarce testing resources in an efficient manner. Another possibility would be to provide individuals with recommendations to limit their social interactions if the algorithm deems them to have a high risk of having been infected. We emphasize that this should only be a *recommendation*. Before sending high-risk individuals into mandatory quarantine, the *estimated* health status would of course still need to be confirmed by a test.

We formulate two desiderata for our algorithm to satisfy, such that we may deem it to be trustworthy.

- **Privacy**. We want to ensure that adversaries who have access to the prediction model cannot infer information on the training data that was used to fit the model. We achieve this goal via the application of differential privacy.

- **Interpretability**. We want to ensure that the model is performing in an expected manner. That is, we want to avoid treating our prediction models as black boxes. We achieve this goal by either building inherently interpretable models or by building black-box models, which we then make interpretable by applying methods from eXplainable Artificial Intelligence (XAI).

Initially, we focus on the first desideratum. We build an inherently interpretable model and induce differential privacy. Concretely, we implement a linear regression based model for the encounter-wise risk prediction task. We ensure differential privacy at this level such that in the next step, when we aggregate the risks of the encounters to the individuals' daily risks, we are still satisfying our privacy goal.

Future work includes the construction of a more complex prediction model, that will require the application of XAI-methods such that we can ensure that our goal of model interpretability is achieved. In this case, satisfying the privacy characteristic will likely be a more challenging task.

For more details, see the full work from Arnold [5].

## 4.10 Natural Language Processing in Activity-based Risk Prediction

The recent outbreak of the coronavirus disease has led researchers and scientists to investigate methods that would help the policymakers take necessary decisions for fighting the pandemic. However, this requires sensitive personal data of the general population, the collection of which may give rise to ethical concerns. This work aims at generating synthetic data to model the spread of disease in an artificial population to perform an activity-based risk prediction. Every person is performing activities in their daily lives to complete some tasks or objectives. Each activity of a person is composed of a sub-element called actions. The machine learning approach for predicting infection risk is based on these actions and their features. One of the most common practices in representing a word as a dense vector is the use of the word embeddings. Repeating the actions of a day, again and again, leads to a resemblance with a text sentence:
*Home - Bus - Work - Bus - Supermarket - Home*

Now, it is desirable to find a way to represent an action in the form of a d-dimensional vector where d can be any positive number. The intuition behind this is to get similar vectors for similar actions. Hence, similar actions would remain close to each other while dissimilar actions will remain further away from each other, in the vector space. In textual data, the words are ordered according to the syntax of the language while in temporal data the concepts are ordered with respect to time. Hence, in order to incorporate the time factors to be able to compute the concept embeddings in a temporal sequence, an embedding technique is needed. Time2Vec [41], a modified Word2Vec skip-gram model which learns embeddings of concepts from temporal data considers the time factor in temporal data. The Time2Vec pipeline has two major components - Vocabulary building and Learning embeddings. The first version of the embeddings reveals that PCA has not been able to separate the observations and therefore, it is more difficult to find any concrete clusters in the embedding data. As a consequence, the embeddings show no clearly discernible pattern. The training was performed several times to generate better embeddings after tuning of the hyperparameters and increasing the number of negative samples that are drawn for every true training sample lead to prominent results.

The embeddings generated after applying all the changes to the Time2Vec model need to be analyzed. k-Means clustering is applied to the 32-dimensional embeddings for the k values ranging from 1 to 9. PCA and t-SNE were both applied to the 32-dimensional embedding data generated by the improved Time2Vec model. PCA was applied to see whether there was a global structure to the embeddings. Then the k-Means clustering is used to color-code the data points in the PCA plot. The four different clusters in k-Means appear to be well clustered in PCA and t-SNE as well. The k-Means clustering on the 32-dimensional embeddings returned a label for every concept in the vocabulary. A label was assigned to every row in the original dataset, based on these cluster labels of the concepts. Then, a classifier is trained using these labels as the target variable. The steps of clustering and classification are kept separate. A Decision Tree classifier with a maximum depth of 3 was trained using the cluster label as the target label which led to an accuracy of nearly 56%. Now, to better understand the propagation of the samples through the tree, the decision rules were extracted in a human-readable format. One of the observations that can be made is that concepts belonging to cluster 3 may have larger group sizes.

To calculate the general statistics of the original dataset, we adopted a statistical approach. We filter the data with the top 10 concepts having the maximum frequency and compute the mean and the standard deviation for two features, group size and duration. We further created a table showing the mean and the standard deviation of these concepts for two features groupSize and duration. We have plotted the means and the standard deviations for group size and duration respectively. By analyzing the plots and the table data, it becomes clear that the group size for the top 10 concepts belonging to cluster 0 is at a slightly higher range than the size for other concepts belonging to the rest of the clusters. Similarly, the value for the duration of the cluster 3 also seems to be at a higher range than the one for the other concepts belonging to the rest of the clusters. We also drew the information that majority of the top 10 concepts in cluster 2 are public transports.

The approach of clustering followed by dimensionality reduction and then color-coding he observations based on cluster labels have revealed some cluster structure within the embeddings. However, we were expecting more fine-grained clusters in the concepts. In this approach, the locations have been used as actions due to the lack of thoroughly detailed information provided by the simulation. The features that were mainly used to weigh the co-occurrences of the concepts are start time, duration, and groupSize. Each of these features can have different values for the same location which needs to be considered. Also, only the facility is not enough to determine infection risk. Based on these facts we conclude that a fine-grained simulation that includes many types of actions performed by the agents in each of these facilities may lead to better results.

In addition to this, a theoretical study to recognize activities and co-activities using sensors has also been presented. This work will allow researchers interested in the domain of activity recognition for pandemics to develop more powerful models in the future. Everyday technology is advancing and the cost of wearable devices is decreasing. These high-end electronic accessories like fitness watches and smartphones, integrated with high-quality sensors have become affordable to most people. Thus, activity logging is becoming popular day by day and is slowly becoming a general practice. Individuals are logging their daily activities like sleeping, cooking, eating, taking steps, etc. Various approaches can identify these activities. We presented the latest research in the field of Human Activity Recognition (HAR). In this thesis, we want to recognize different activities considering the epidemic spread in mind. As discussed in a previous section, activities of people are one of the key factors which govern the epidemic spread, and hence, we need to identify and cluster activities that involve a lot of interaction and contact between people. We propose a theoretical approach in this section consisting of the following three steps: Individual activity recognition, Bluetooth as a social feature and Interaction geometry and Co-activity detection.

For more details, see the full work from Ghosh [28].

### 4.11 Summary of technical findings

A big part of the technical side of the project was the development and testing of the simulation framework. This encounter-based simulation was ultimately chosen, since it enables a very fine-grained look on the disease spread in the overall population beyond basic statistics, modelling social encounters and a more natural disease spread, according to activity and location. In turn, the experiments done on top of this simulation could be done with a variety of different scopes and objectives, looking at such scenarios as individual risk-prediction, location based disease spread, privacy preserving Machine Learning, explainability methods for disease spread prediction to facilitate trust-building or testing the effects of methods for pandemic control that are in turn informed by Machine Learning predictions.

Among the central key findings for the technical part are the fact that simulating data is a recommended process for the development of this type of model, where the actual data to be used is scarcely available ahead of time. Not only was it possible to test different methods and strategies with the simulation data, but the data also allows for more informed judgements on the necessity of collecting a certain type of data for deployment in a real use-case. The approach for this was led by a minimal invasive principle, only collecting the data that is most necessary without a significant drop in predictive power.

Similarly experiments have shown that it is possible to design such Machine Learning based prediction systems that adhere to a high standard in terms of privacy and explainability, without sacrificing too much accuracy. This is especially relevant for the ethical deployment of such models, which are often forced to collect sensitive personal data from individuals.

The final key-finding is about the necessity of enough coverage in the data collection process to use a system for targeted advice in terms of risk assessment. Unfortunately, systems require widespread adoption of data collection. The more fragmented the data collection, naturally the performance decreases, however experiments have shown that this adherence needs to be much higher than it might previously be believed, to still be able to offer good predictive power. Similarly, compliance with imposed measures to combat disease spreads was required to be even higher, which suggests that measures would need to either be accepted voluntary by the vast majority of people or that there are only measures in place, which are easy to enforce.

The following part will offer a broader summary of the conclusions of this project and offer further insights of the interplay between the technical results and the ethical and legal considerations for the health scenario.

# 5 Conclusions and Recommendations

In this interdisciplinary study the intersections between ethical considerations, legal implications, and the use of sophisticated machine learning techniques were traversed. This exempted the project from the typical pitfalls of siloed thinking where ethical AI is discussed by the philosophers, legal AI by the lawyers and the machine learning left to the data scientists. The main learning of this project is that the ethical and legal aspects of AI is not a post factum checklist and compliance exercise. It is an essential part of the design and execution of the particular solution to a clearly defined use case.

## 5.1 Values by design

Values by Design is used as part of the design specifications for this project, and also as a guide to assess the AI model and its application in fighting health pandemics by strengthening the human decision making. This means that a process-based approach should be followed, namely that the principles apply to:

- The acquisition of data.
- Design of the AI model in different scenarios.
- Implementation of the AI.
- Impact of the use of the AI.

It is an iterative process in which the multi-disciplinarity of the approach is emphasized and used as an asset to develop a clear ethical and legal framework specifically for the health context, as well as a relevant and reliable AI model that could contribute to improved decision making.

## 5.2 Design of an ethical and legal framework for AI in fighting pandemics

### 5.2.1 Ethical perspective

Various ethical theories should act as lenses in the development of an ethical framework, as no single theory on its own will allow for the sensing of the context of an ethical use case, and for the pre-assessment of the impact on core human rights. This assessment will lead to the careful weighting of the respective consequences, and the description of inevitable tradeoffs, which should be optimised for the maximum benefit to all stakeholders involved.

The different lenses (or ethical theories) are:

- The Rights Perspective
- The Justice or Fairness Perspective
- The Utilitarian Perspective
- The Common Good Perspective
- The Virtue Ethics Perspective.

Ethical user stories were used to assist in the development of the Machine Learning models. In other words, each model was described in terms of its user, its objectives or functionality and its outcomes. Acceptability criteria was defined according to the ethical lenses of privacy, justice, transaparency and reliability.

### 5.2.2 Legal framework

A human centred approach should be followed in developing ethical AI models for managing health pandemics. It is concluded that the following key legal principles, which are common to most international policy documents on ethical AI, apply, namely:

- Transparency
- Accountability (auditability)
- Dignity, fairness, equality
- Privacy and data governance
- Other human rights considerations, e.g. human autonomy and oversight.

### 5.3 Findings on Privacy, Fairness and Explainability

Although the key legal principles indicated above apply in general to the development of ethical AI models for managing health pandemics, this project produced some interesting findings regarding privacy, fairness and explainability (transparency). In the development of the disease spread simulation three elements or stages were identified (see detail technical discussion above), namely mobility and encounters, long and short term social networks, and infections or disease spread. In the encounter-based risk prediction individual risk prediction is possible and made use of differentiated privacy, which means that no outside person can get access to the individual data which is then kept private.

This encounter-based prediction also allows for explainability - the various features of the AI model are described to assist eventually in better decision-making in government. It is important to note that there is also a spatio-temporal element in this model since encounters are related to time and space. The results of such a spatio-temporal simulation assist decision-makers to regulate access to different facilities at specific times, and individual citizens who can decide better how and when to avoid high-risk facilities.

Activity-based risk prediction combined with the spatio-temporal element provides a more comprehensive risk prediction model that allows for what-if scenarios.

Fairness is accommodated in this AI model by ensuring the data simulation is done fairly and procedural fairness is applied throughout the AI life cycle. Outcome fairness can be achieved if the implementation of the AI model does not cause harm, discriminatory or inequitable impacts.

### 5.4 The big trade-off

Situational ethics is the position that moral decision making is contextual or dependent on a set of circumstances. This has become quite clear during the course of this interdisciplinary study. The prior assessment of the probability and impact of unintended consequences in the use of models with agency is critical for the selection of mitigation strategies. There is no one-size-fits-all set of criteria and checklists which will safeguard us once and for all against ethical and legal failures. It is as much of an art as it is a process to carefully weigh the personal with the public interest, the right to life with the right to privacy and the advantages of sophisticated technology with the need of transparency. The set of ethical and legal lenses presented in this study, may at the very least provide the tool-set, perspectives and parameters to get the discussion going, and to integrate the ethical and legal perspectives to permeate the design of AI systems.

## References

[1] AlgorithmWatch. Ai ethics guidelines global inventory, 2021.

[2] M. Ananny and K. Crawford. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*, page 1, 2016.

[3] M. Ananny and K. Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, 20(3):973–989, 2018.

[4] M. Anschütz. Simulation and mining of social networks to predict individual infection risks, 2021. advised and supervised by Edoardo Mosca, Tobias Eder, and Georg Groh.

[5] J. Arnold. Explainable and differentially private machine learning models for pandemic control, 2022. advised and supervised by Edoardo Mosca, Tobias Eder, and Georg Groh.

[6] E. E. Avery Bossert Clark. Policy implications of models of the spread of coronavirus: Perspectives and opportunities for economists. *National Bureau of Economic Research*, 2020.

[7] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

[8] G. Bezirganyan. Ethical ai for pandemic control based on spatio-temporal data, 2022. advised and supervised by Edoardo Mosca, Tobias Eder, and Georg Groh.

[9] BFDT. Bayerisches forschungsinstitut für digitale transformation, 2020. bidt.digital. [Online.

[10] P. Boddington. *Towards a code of ethics for artificial intelligence*. Springer, 2017.

[11] D. Brand. Algorithmic decision-making and the law. *JeDEM-eJournal of eDemocracy and Open Government*, 12(1):115–132, 2020.

[12] C. Busch. *Algorithmic Accountability*. Bundesministerium für Bildung und Forschung, Berlin, 2018.

[13] R. Chatila and J. C. Havens. The ieee global initiative on ethics of autonomous and intelligent systems. In *Robotics and well-being*, pages 11–16. Springer, 2019.

[14] K. Chen, J. See. Artificial intelligence for covid-19: Rapid review. *Journal of Medical Internet Research*, 22(10), 2020.

[15] E. Commission. Ethics guidelines for trustworthy ai, 2019.

[16] E. C. C. Committee. Council of europe consultative committee of convention 108, 2019, 2019.

[17] E. Council. Council of europe conventions, 2018.

[18] P. De Vos, W. Freedman, and D. Brand. *South African constitutional law in context*. Oxford University Press, 2014.

[19] B. des Innenn und für Heimat. Gutachten der datenethikkommission, 2019.

[20] V. Dignum. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature, 2019.

[21] D. Drothler. Realistic pandemic spread simulation to enable machine-learning-driven prevention, 2021. advised and supervised by Edoardo Mosca, Tobias Eder, and Georg Groh.

[22] L. Edwards and M. Veale. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, 16:18, 2017.

[23] D. Fang, Y. Zhang, and W. Wang. Complex behaviors of epidemic model with nonlinear rewiring rate. *Complexity*, 2020, 2020.

[24] F.A.T.M.L. Fairness, accountability and transparency in machine learning, 2016.

[25] T. Fischer. Testing strategies fro pandemic control, 2021. advised and supervised by Edoardo Mosca, Tobias Eder, and Georg Groh.

[26] W. E. Forum. World economic forum, 2019.

[27] W. W. W. Foundation. World wide web foundation, 2017.

[28] T. Ghosh. Ethical ai and natural language processing in activity-based risk prediction for pandemic control, 2021. advised and supervised by Edoardo Mosca, Tobias Eder, and Georg Groh.

[29] A. Group et al. From principles to practice—an interdisciplinary framework to operationalise ai ethics (p. 56). vde/bertelsmann stiftung, 2020.

[30] R. Guerrero. A framework for ethical decision making in problem definition & project selection, 2019.

[31] S. Hallensleben and C. Hustedt. *From principles to practice: An interdisciplinary framework to operationalise AI ethics.* Bertelsmann Stiftung, 2020.

[32] N. Harding, R. Spinney, and M. Prokopenko. Phase transitions in spatial connectivity during influenza pandemics. *Entropy*, 22(2):133, 2020.

[33] I.C.D.P.P.C. International conference of data protection and privacy commissioners, 2018.

[34] A. G. Kuperman M. Small world effect in an epidemiological model. *Physical Review Letters*, 86(13):2909, 2001.

[35] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak. Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review. *Chaos, Solitons & Fractals*, 139:110059, 2020.

[36] S.-Y. Lee. Decision intelligence and machine learning for pandemic management, 2022. advised and supervised by Ben Herbst, McElory Hoffmann, and Georg Groh.

[37] D. Leslie. *Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector.* The Alan Turing Institute, London, 2019.

[38] D. Leslie. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of ai systems in the public sector. *Available at SSRN 3403301*, 2019.

[39] M. A. Lopez. Ethical machine learning for pandemic control, 2021. advised and supervised by Edoardo Mosca, Tobias Eder, and Georg Groh.

[40] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[41] P. M. Dynamic network models and graphon estimation. *Annals of Statistics*, 47(4):2378–403, 2019.

[42] B. McCall. Covid-19 and artificial intelligence: protecting health-care workers and curbing the spread. *The Lancet Digital Health*, 2(4):e166–e167, 2020.

[43] L. Mitrou. Data protection, artificial intelligence and cognitive services: Is the general data protection regulation (gdpr)'artificial intelligence-proof'? *Artificial Intelligence and Cognitive Services: Is the General Data Protection Regulation (GDPR)'Artificial Intelligence-Proof*, 2018.

[44] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679, 2016.

[45] C. Molnar. *Interpretable machine learning.* Lulu. com, 2020.

[46] F. Morina. Decentralized machine learning for pandemic control, 2021. advised and supervised by Edoardo Mosca, Tobias Eder, and Georg Groh.

[47] U. Nations. United nations universal declaration of human rights, 1948.

[48] M. E. Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

[49] O.E.C.D. *Recommendation of the Council on Artificial Intelligence.* OECD, OECD/LEGAL/0449, Paris, 2020.

[50] H. L. E. G. on Artificial Intelligence. European commission. *High Level Expert Group on Artificial Intelligence*, 2019.

[51] N. Phan, X. Wu, H. Hu, and D. Dou. Adaptive laplace mechanism: Differential privacy preservation in deep learning. In *2017 IEEE international conference on data mining (ICDM)*, pages 385–394. IEEE, 2017.

[52] T. Philbeck, N. Davis, and A. Larsen. World economic forum, 2018.

[53] I. Price and W. Nicholson. Artificial intelligence in health care: applications and legal issues. *SciTech Lawyer*, 2017.

[54] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[55] E. Schwella. *South African Governance.* Oxford University Press, Cape Town, 1 edition, 2015.

[56] P. D. P. C. Singapore. Personal data protection commission singapore, 2020.

[57] R. South Africa. *Constitution of the Republic of South Africa.* Government Printing Works, Pretoria, 1996.

[58] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo, and L. Floridi. The ethics of algorithms: key problems and solutions. *AI & SOCIETY*, pages 1–16, 2021.

[59] E. C. D. P. Unit. Digital solutions to fight covid-19, 2020.

[60] S. Vallor, B. Green, and I. Raicu. Ethics in technology practice. *The Markkula Center for Applied Ethics at Santa Clara University. https*, 2018.

[61] I. van de Poel. Embedding values in artificial intelligence (ai) systems. *Minds and Machines*, 30(3):385–409, 2020.

[62] P. Voigt. The covid-19 guidelines of the edpb, 2020.

[63] J. Wang, X. Kong, F. Xia, and L. Sun. Urban human mobility: Data-driven modeling and prediction. *Acm Sigkdd Explorations Newsletter*, 21(1):1–19, 2019.

[64] M. Wu. Activity-based risk prediction for pandemic management with ethical ai, 2022. advised and supervised by Edoardo Mosca, Tobias Eder, and Georg Groh.

[65] A. Xiang and I. D. Raji. On the legal compatibility of fairness definitions. *arXiv preprint arXiv:1912.00761*, 2019.

[66] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[67] ZACC. Prinsloo v. van der linde and another, 1997.

[68] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett. Functional mechanism: regression analysis under differential privacy. *arXiv preprint arXiv:1208.0219*, 2012.

[69] S. Zhang, H. Tong, J. Xu, and R. Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019.