
# BRIEFING NOTES

# BN-75-Emerging technology and military application-Aug2021

## AI PUBLIC POLICY, ACCOUNTABILITY, PRIVACY, IMPACT ASSESSMENT, AND EXPLAINABILITY

Authors: : Mohammadreza Nematollahi[1] and Kash Khorasani[2]

1 Graduate student, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

2 Professor, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

## ABSTRACT

In this report, frameworks for further studies in the field of "public policy and artificial intelligence" have been investigated. We mainly focused on the currently raised concerns regarding how integrating artificially intelligent systems with human society impacts our lives and society. More specifically, criminal liabilities of different stakeholders, a framework for holding the service provider agencies accountable concerning the social good, and implications on privacy need to be addressed. Compared with conventional design failure lawsuits, it turned out that behavioral approach for determining the liabilities fits better the context of AI technologies, yet, some further clarifications are required. For active and ongoing management of associated risks with these technologies, the famous algorithm impact assessment framework seems to be a good fit. At the same time, the role of explanation generation, which is another way for keeping people and service providers utilizing artificial intelligence technologies, in touch has been studied. One also requires to regulate the direct and indirect implications of these technologies on privacy due to their capacity to deanonymize the protected information or their implication on the data markets. Certain efforts and suggestions regarding privacy issues are evaluated. Finally, a general impact assessment framework is adopted based on which one can analyze the risks and benefits of applying these technologies to particular problems.


## INTRODUCTION

Based on estimations of Cisco systems, between 2008 and 2009, ratio of things or objects connected to the Internet over people connected to the Internet became more than one, and a new type of systems, called Internet of Things (IoT), was born [1]. IoT is a system of interlinked objects, animals, people, computing devices, digital systems, and mechanical devices that can communicate to each other and are able to form a network of devices [1]. This concept was described as packets of data that are transferred between different nodes, which these nodes can be home appliances or sections of a factory, such that these nodes are integrated as an automated system [2].

From the beginning of the past decade, artificial intelligence (AI) systems started becoming ubiquitous in all science and technology branches. Many AI ideas reached commercial standards and are available in the market, while there are tons of other ideas that are passing their way toward maturity and will be available within the next few years. A full chapter in [1], from Organization for Economic Cooperation and Development (OECD), is dedicated to surveying AI's contribution in some relevant application fields such as finance, transportation, agriculture, marketing and advertising, health, science, security, and even criminal prosecution. More familiar

readers can also refer to [2] [3] [4], in which authors addressed some AI applications in video processing, self-driving cars, and even self-driving airplanes, and also [5], which studied application of AI-facial recognition systems for criminality evaluations.

AI solutions are very desirable for tech companies and even governmental associations, and this has led to a rush for unregulated integration of those systems with human society and created a very competitive arena of science and technology from the economic and political points of view, even at the international level. However, like other new technologies, AI technologies also have good or bad impressions on society and people. In a well-regulated environment, technologies can help society solves its problems and make it a better place for people to live, while careless integration without being prepared enough, can make the situation worse. This is while we have witnessed some failures and problems in applying AI technologies in close contact with human society during the past few years. These examples brought to facade the limitations of the AI technologies and the current regulations' inefficiencies in dealing with them [6] [7] [8].

In [6] the authors addressed the case of Uber's self-driving car crashed a pedestrian in 2018. Although the main problem at that moment was due to lack of active emergency brakes and the AI algorithms had nothing to do with this failure, as revealed later, the software also was not appropriately designed to detect the pedestrians out of the crosswalk, implying that Uber's self-driving technology had not considered the human interactions in its design [6] [9] correctly. Although it was not intentional, yet it showed that, in fact, detecting even such apparent deficiencies in AI models is not an easy task since they usually are complex models beyond human capabilities to justify them. The case of Amazon's automatic recruiting system has been addressed in [7].

As of 2015, they realized that their system is not gender-neutral and excludes women from the candidate pool by down-scoring people who have the word "woman" on their resume. It came out that the main reason behind such behavior was, in fact, the dataset that had been used for training the AI ranking system. Some articles as in [10] conduct studies regarding the main factors causing such discriminations, which can be repeated in the AI technologies without choosing carefully the underlying datasets. The case of Facebook photo-tag suggest system [8], although is not an example of explicit failures, raises alarms for privacy issues, where [11] addressed another case of privacy ignorance in this field through the famous case of Facebook's Cambridge Analytica data scandal, which raised the international concerns regarding privacy.

Considering AI engines along with big data analytics tools and the internet of things (IoT) provide a highly proficient computational system, which can solve many real-world problems, for which previously there were no computationally efficient analytic solutions. Compared with traditional analytic design solutions, in the case of AI systems, the designer, instead of investigating the related concepts and models deeply, need only to provide an informative dataset and a sufficiently complex neural network structure, which most of the time will be chosen based on a trial and error process. The system investigates the dataset and learns the knowledge structure by itself in an autonomous fashion following a prespecified optimization strategy.

Big data analytics tools and IoT infrastructure help designers handle the data acquisition steps, while neural networks and their associated learning algorithms handle the rest. At the same time, the system may also be equipped with some algorithms called reinforcement learning for facing and investigating unknown environments. These capabilities provide AI technologies a unique characteristic that distinguish them from other technologies and resembles human thinking capabilities. They can make an inference, learn, and discover, all of which provide them a level of autonomy in decision-making and lead to a human-like behavior, enabling them to go beyond what they have been designed for, initially.

Moreover, in contrast to other systems, a designer only needs to decide among some high-level options during the design procedure to provide the required datasets for training; after that, the system uses the provided dataset to tune the parameters to reach the mathematically described performance criteria. In most cases, the design procedure is not intuitive, and the designer comes up with a large number of parameters specified by choice of the preferred high-level structures. Therefore, even for experts in the field, it would be very hard or even impossible to probe and explain the system's behavior after design; therefore, AI systems have usually been considered black boxes [12], which means, we only have very high-level information about their behavior.

Data is the most crucial factor in developing AI technologies, and consequently, AI increases the value of data much more than before. This makes companies even less eager to protect the users' data and privacy [11]. This is while, nowadays, human life has been highly tied to social networks and the internet, and due to the “strong effects of social networks” [8], people recklessly share images, videos, and other types of information related to every aspect of their lives through these networks. At the same time, they use the internet almost everywhere as the primary source of information or surf the web for its fun. They do so, even without a proper

understanding of how others can utilize this generated information and what is going on behind the scene.

Social networks such as Facebook, the biggest among them [13], collect a vast amount of data through surfing the user profiles and information they provide by sharing their image and videos along with tagged information when websites and search engines like Google [14] and Amazon [15] can also track the user's search history and other related information to gather more information about the users. The primary goal of these data acquisitions is to personalize the services provided to the users, yet, there is no guarantee that those data will not be misused in the future intentionally or unintentionally due to the capabilities of the AI, while also there is no strong way to hold them accountable and liable even concerning their own published ethical guidelines and accountability agreements.

This is the law of nature that everything will grow better where the conditions are more favorable to them, and AI technologies are not an exception. Therefore, as we expect, many AI ideas, especially those AI facial recognition systems, have been developed and implemented over the social network and internet infrastructure, and their destinies are intertwined. We cannot regulate the AI without holding those tech companies behind social networks and the internet accountable, and of course, any regulation will also affect how those tech giants interact with users.

This is while, as statistics show [16], societies are becoming more dependent on the services the companies provide, and the companies reached international influences as well. Besides, we should always have this on the side of our mind that we are yet in the development phase of AI technology, a technology proven to have substantial economic and geopolitical implications, and there is an international rush towards it. Hence, a better solution should take into account these factors as well.

This report is mainly concerned with the frameworks for addressing criminal liabilities, accountabilities, and privacy issues related to AI technologies. Many debates exist among researchers and there are many research publications [17] [18] [19] worldwide addressing those questions and have tried to propose optimal solutions for how AI systems should be integrated into society, how and to which level governments should intervene in the process to ensure the overall net benefit to the society, both in the long-term and short-term, while also taking into account the current state of the technology in its development phase, its progress rate, the dependency of society on those technologies, and economical and geopolitical factors.

Lawmakers should adequately address the legal liabilities of the designers and agencies using AI systems. This should be addressed as soon as possible since any ambiguity impedes the path towards the next generations of AI systems and stifles the investments in this field [20]. Any regulations, minimally, should ensure a smooth path in testing and rolling out at least limited versions of the technology; otherwise, one must not expect that companies will be eager to invest in the fields since they are not safe areas of investment and are highly risky.

What if the law forces these companies to recall all the products and ask for redressing due to some minor failures? Conventionally, design defect lawsuits are applied in the cases of technology failures. However, in new AI technologies, those lawsuits concern the AI companies the most and seem unsuitable and are very restrictive for new tech companies, especially those smaller ones. In these lawsuits, no matter who wins the trial, they would be costly for AI developers due to the AI systems' unique characteristics, which distinguish them from other analytic solutions.

AI systems can also be part of automated decision-making processes, for example, risk assessment tools in courts [21] or any other automated decision-making processes utilized by other agencies. These systems are products under the control of agencies providing services to the public. In such circumstances, proposing mechanisms to hold those agencies accountable for social norms and fundamental democratic freedoms of people and monitoring those responsibilities has a vital place in AI public policy discussions and research [22], [23]. Nothing should prevent agencies from fulfilling their responsibilities to protect fundamental democratic values. At the same time, the complexity of the underlying processes is not a good reason to run away from the responsibilities, yet, we need appropriately specified guidelines for holding them accountable and preventing them from hiding themselves behind the black-box characteristics of the AI systems.

Requirement of accountability go beyond those criminal liabilities and are in the scope of the agencies' social and ethical responsibilities, even when the law may be silent regarding those issues. For example, right now in most of the countries, there is no law to force agencies to disclose their decision-making algorithms, and there is no well-defined mechanism for monitoring those agencies or holding them accountable regarding fundamental democratic freedoms of the people in society, and social norms [24]. These problems should be addressed in the early development stage of the AI systems. It is worth mentioning that there is a direct relationship between an agency's accountability and its trustworthiness by society, where an

agency's social trustworthiness is a consequence of respecting their responsibilities regarding society.

Without a well-defined framework for monitoring and holding the public agencies accountable and having an explicit agreement of their responsibilities in the society, people and agencies will lose touch, make it impossible to detect unfairness or any other problem in the decision-making process and making them unable to object, which consequently makes them hesitant regarding those agencies, which implies having a properly defined accountability framework benefits both sides.

Privacy is a long-lasting human right in free and democratic societies that should be revised and adapted based on the recent changes in technology trends and the changes in the interaction among people and the products and services. Traditional privacy and data protection laws have lost ground due to the increasing dependence of the people on the internet and other software connected to the internet. The traditional laws depend mainly on consent agreement contracts between people and service providers and attribute more responsibility to people to protect their privacy by carefully reading the agreements and finding alternatives. In contrast, in the era of AI systems, due to dependencies of AI systems to having informative enough datasets, designers become highly dependent on the data, and therefore the data protection laws should put more responsibility on their shoulder, while as also discussed in [25], this does not mean at all that people should not care about it.

Data for AI systems is like food and energy for humans, without which we cannot expect what we currently expect for the future of this technology course. Hence, current data protection and privacy laws, which were mainly based on not disclosing the data, are not compatible with these new technologies, and revising and reforming those laws and methodologies are required. On the other hand, due to their speed and capabilities in processing large amounts of data and their complex internal structure, AI systems can highly affect people' privacy. The problem is not just about AI system capabilities in deanonymizing the protected information and mining the complementary data from large data sets [26], but AI systems commonly have also been used as decision-making service providers, and there, due to their highly complex structure, one cannot ensure whether sensitive information has been used and if yes, have they been used correctly? This also highlights the role of explainability in achieving safer and more privacy-aware technology, which minimizes the risk and harms to people and society [27].

Following our research path, the recent facial recognition impact assessment (FIA) framework [28] has also been reviewed, which although is mainly developed for AI-based facial recognitions systems, it has a great potential and can be applied for other applications as well. This framework, along with the well-known algorithm impact assessment, which considered the accountability of agencies using AI algorithms and autonomous decision-making processes, and also beside the law of General Data Protection and Regulation (GDPR) [29], can serve as a more or less complete framework for studying different aspects of interactions of AI technologies and society.

## CRIMINAL LIABILITIES

The first step toward regulating the application of AI technologies in society should be to clarify the role of law, which is the most significant player. Hence, in this section, we will discuss and present a proposed approach toward this problem, trying to clarify the criminal liabilities of AI systems, designers, and service providers. Policymakers and lawmakers should be careful about the economic and geopolitical factors in proposing regulations and supporting laws, to not damage the small AI-oriented business; because of a few problems, one should not underestimate their capabilities. At the same time also considering the global economy and geopolitics rush in becoming the AI-leader of the world, the significant impact of law should be addressed as soon as possible since any ambiguity can impede the path toward the next generation of AI systems, and imply losing a potentially significant flow of investments.

AI systems can be separated into two main parts, AI software, which serves as the system brain, and physical tools, designed based on the context. Any problem in physical tools can be handled like other technological failures, based on the currently available legal supports; hence, these problems are not the subject of this study. Regarding the system's AI brain in order to apply the design lawsuits, one can only probe millions of lines codes, massive datasets, millions of parameters, and the sensors and actuators' compatibility to justify only high-level design parameters under the designers' control, while in most cases, they are not intuitively related to the underlying context, and doing so, is like checking all the brain and other body cells in a criminal's body in a court trial. One also can ask, who are the right persons to do that? Who should support them in doing so?

Furthermore, regarding the punishments and compensations, holding designers responsible and only blaming them in all cases is like punishing a criminal's parents and teachers, while the criminal himself/herself had enough autonomy in committing the crime. It should be noted that design defect lawsuits usually force the companies to recall all the products to fix them and redress, because in their logic, it will be very likely to lead to the same problem again, which

means a substantial financial loss for the companies, where in some countries the companies are also subject to further punishments due to designing and distributing defective products. This is while some of the AI systems can adapt their behavior as time pass, and at the same time like any other software, one can solve their problems through applying patches remotely by issuing updates online.

One needs to properly define and adapt the criminal liability models to cover the context of AI systems. In [30], Gabriel Hallevy suggested treating the AI systems in the same way as humans committing a crime. In this sense, instead of internal structure, behaviors are subject to law, which better fits the case of AI systems in criminal conduct. We will also follow this human-like treatment in this report based on what has been discussed in [30], [31] and references therein as a framework of AI system criminal liabilities.

For criminal liability to be shown, as Hallevy discussed, firstly, one needs an "*actus reus*" to exists, which could be either direct offensive action, failure to act appropriately, or omission. The existence of an *actus reus* is a necessary condition, while in the case of strict liability offenses, for which no "*mens rea*" (mental capability) is required to be held liable, it would be sufficient as well. For example, in certain jurisdictions, having some kinds of drugs will result in criminal liability, regardless of whether the defendant knows that he/she has the drugs or not [32]. In other cases, *mens rea* is also required. The lowest level of *mens rea* required for some cases is negligence, which is applicable where a reasonable person should know he/she may commit a crime. In other cases, intention for committing the crime is also a requirement, and finally, the highest level of *mens rea* requires the availability of knowledge to commit the crime.

In contrast with other technologies, AI systems fit into the class of entities, that can have both requirements simultaneously, and hence, they are subject to criminal laws even by themselves! Three liability models applicable to the case of AI systems have been suggested by Hallevy [30] and will be reviewed here for completeness. Based on these models, one can attribute the liabilities to either users, developers, or AI systems proportionally.

**Perpetrator-by-another:** This is the same model that is used in cases where the crime will be committed by mentally deficient persons such as children. The offender (in this case, AI systems) is innocent even in case of strict liabilities and is not aware of the possibility of committing a crime due to lack of enough mental capacity. In such cases, who causes the innocent entity to commit the crime, is criminally liable by the perpetrator-via-another model, and the innocent entity will be considered an instrument for him/her to commit the crime. One needs to determine

the perpetrator-via-another liability based on his/her mental state, while the system's actions in committing the crime will be attributed to him/her. In the case of an AI system, the perpetrator-via-another can be either the developer or the user, and assuming that the user could not program the AI system and only can give it commands from an available set of commands, it would be reasonably easy to determine which one should be held liable.

**Natural-probable-consequence:** Another related scenario is when users and developers have neither criminal intent nor the knowledge that it may happen and are in a negligent mental state. AI system, during the execution of its daily tasks, may commit a crime. Although the user and developer did not know the possibility of the outcomes until it came out that the system committed a crime during the tasks' execution, and they also had no previous plan in doing so and did not accompany the AI system during the process, the users and developers are liable since the crime is the natural consequence of their actions.

This model will also apply to cases that require different state of mind for the user and developer; they are not the main perpetrator of the offense but are the intellectual perpetrators. These two different scenarios will be handled differently, as the user or the programmer in the former scenario lacks the criminal intent, but in the latter, they had such intentions and will be considered as accomplices. The third scenario is a combination of both previous ones, in which during the committing the first crime in which the user or the developer accompanied the system for it, it may commit another crime either instead or in addition to the first crime beyond the accomplices' knowledge.

**Direct Liability of the AI systems:** Hallevy's article [30] also addressed the cases where the AI systems are directly liable. To hold an AI entity liable by law, one needs to determine the existence of the requirements, *actus reus* as the necessary condition, and *mens rea*, if applicable, based on the context. Usually, attributing the required *mens rea* to AI systems is much more challenging. AI system may resemble most human cognitive capabilities and creativity even in a more efficient form, but those are neither sufficient to prove nor exempting from direct criminal liability of the entity.

To hold an entity liable, one needs capabilities to acquire the related form of knowledge, generate criminal intent or negligence, or other mentally related capacities. The knowledge acquisition capability of the AI entity should be related to the criminal context. It should contain the minimum number of required sensory systems and required inference capabilities, which in most cases these capabilities well resemble the human counterpart regarding the intent

generation. The AI systems may lack some major capabilities of humans, and hence it is not an easy and trivial task to attribute the AI systems, a specific criminal intent or negligence. At the same time, there is no reason to not being able to do so; hence a more detailed analysis is required.

Towards this end, Hallevy also compared the case of AI systems with those classes of humans which are exempted from being liable with respect to criminal laws and concluded that in cases where the criminal liability's primary requirements exist in an AI entity, none of those exemptions are applicable.

In cases where humans are liable, determining the punishments is trivial and is based on the currently available methodologies and is not subject to this study. However, when one holds an AI system liable, what are the punishments? Hallevy believes punishments are adaptable to the case of AI systems, in the same way as we have adaptations of the human criminal punishments for liable agencies or companies, and proposed some of the possible adaptations.

However, we believe the proposed model has some major deficiencies since, in contrast with human criminals, which will be considered free creatures responsible for their own faith, the AI systems are machines owned by the users. Hence, applying some punishments are not relevant in this case since they may result in financial losses to third parties. Simultaneously, although the quality of cognitive properties is not a matter in determining liabilities and only having the required capabilities are sufficient, they are one of the parameters that should be considered in defining proportional punishments, such that it effectively prevents future offenses while it does not cause others harm.

Going from design failure lawsuits to Hallevy's model of criminal liabilities is quite a significant step. However, Hallevy's article looked at the subject mostly from a lawmaker point of view based on a general understanding of AI systems and lacked the capability to handle financial or technical issues. One such problem has been addressed in [20], and is related to the cybersecurity of AI systems; what if a software virus causes an AI system to commit a crime?

Hallevy also did not address the required steps in implementing this framework, which should be addressed by policymakers in a way that it guarantees a smooth path towards the maturity of both the liability framework and AI technologies. In fact, addressing the criminal liabilities alone cannot ensure the safe interaction of AI systems with human society. Hallevy's article is open concerning some technical standards required by the definition of safe operation of AI systems

and defining developers' and users' negligence, and as the first step, one needs to properly define the required standards of operation and training process of AI systems. This step itself requires one to provide an AI friendly environment in which developers could release the early versions of their products for real-world experiments, by which policymakers can facilitate the path in training AI systems, since in order to achieve the required level of standards, developers should have access to informative enough datasets, which are otherwise only accessible to some limited number of big tech companies.

## ACCOUNTABILITY OF AGENCIES USING AI SYSTEMS

The requirement of accountability framework goes beyond those criminal liabilities and are in the scope of agencies' social and ethical responsibilities, even when the law may be silent regarding those issues. Therefore, accountability frameworks should be part of the required standards for both designing and using AI systems.

Algorithm Impact Assessment (AIA) [24] defines procedures to ensure agencies' accountability scope while considering their unwillingness to disclose their decision-making algorithm publicly. Generating explanation is another method to keep the service providers and people affected by the system in touch, as addressed in [33]. The articles [24] and [33] provided extensive discussions around the two topics of accountability and generating explanation, which will be discussed subsequently as part of the suggested framework for AI policymakers.

### *A. Algorithm impact assessment:*

Algorithm impact assessment (AIA) as discussed in [24] is a framework for agencies using autonomous decision making systems to assess their automated decision-making processes and ensure the accountability scopes of each process and provides a mechanism, based on which society can monitor the agencies servicing the society with autonomous decision-making systems such as AI assistive decision-making systems. If the process is appropriately standardized, it may also impact even the developers by leading to fair competition among them.

An implementation of the AIA should include all the autonomous decision-making processes if they have a signature in the publicly available definition of the agency's autonomous decision-making processes and it consists of the following steps:

**Publishing the definition of their autonomous decision-making process:** It provides a general (not a technical) overview of how the system accomplishes its goals.

It would be better to have a slightly broader definition, which covers not only the software aspects but also human and social factors involving the process and also short history of biases in the underlying context. Yet, the definition should not be trivially broad since it will overload the agency without gaining anything significant. In contrast, a narrow definition negatively affects the scope of the agencies' trustworthiness as it cannot cover an essential aspect of the process. The definition should cover the essential aspects of the data collection, training, and evaluation phase since these are very important in the context of AI system. The definition is not necessary to be solidly defined; agencies can revisit, refine, and extend the definition as time passes to increase the public trust or update the application policies or improve the solution.

**Notifying the public about how the system will influence them:** The usage policies, potential impacts, and initial aims of the system should be published and be accessible to the public. This requirement reflects where the technologies will be used and where accountability researchers and communities' advocates should focus. This is also important since it is closely related to user intentions of utilization.

**The agencies should be experts of their system:** The agencies should have enough capability to understand their decision-making systems' influences and evaluate how this system will impact different communities. They should have a plan for managing the potential negative impacts if any exists or may arise during deployment. Their evaluation should be detailed enough to be useful for external researchers, while at the same time, they should provide a non-technical report for the general public as well. In cases where there may be a potentially negative impact for some communities, the agencies could proactively engage those communities during the assessment process to ensure their concerns and convince them that the system will have a net benefit for them.

Agencies should be concerned with both the allocative harms and representative harms. The first one is related to those impacts, which may deprive a community of accessing something, and the second one is mainly about issues related to the position of a community in society. The main challenge is that, even while studying the potential impacts and during the evaluation phase, they need to have a sense of rights or wrongs, which varies between communities and from place to place severely.

**Providing reasonable ongoing accessibilities:** They should provide a proper level of ongoing accessibility for external researchers to further examine the system in practice once the system

is deployed. It is also a good practice to give those affected communities a chance to introduce their researchers to properly examine the system. This process should be ongoing as the context itself will change, and attitudes will vary as time passes. The level of accessibility depends on the system and its purposes, and will mainly include:

- Provide access to input and outputs of the system for evaluation and provide a relevant and straightforward description of the decision-making algorithm (not the process of decision making, which is mostly technical).
- Provide access to the training data and record of past decisions.
- Agencies should also openly publish the system's access-log and their joint research to gain the people' trust.

The agencies should assess their automated decision-making processes, no matter how they acquire the system, before deploying them. Sometimes, these steps lead to conflicts with trade secrecy. However, as agencies will bear all the responsibilities, they should ask vendors to transfer all the system's rights to them. Regarding the implementation, they can either go through all the processes and publish the documents once, or go through it step by step while also having a public noticing and commenting process to adapt the essential parts.

### *B. The role of explanation:*

Having explanation for decisions made by the automated decision-making systems keeps the society and the agencies in touch and highly improves the agencies trustworthiness. Authors in [33] have tackled the role of the explanation in holding the public agencies and AI-service provides accountable. An explanation is a human interpretable description of the decision-making, by which a set of inputs result in particular outputs. This is supposed to justify the outcomes instead of describing the internal processes that have led to those decisions. An explanation provides the answer to the following questions:

- What are the primary and determining variables in this process? Answering this question enables the observer to see if the right variables have been used.
- If some of the variables can be used both in the right or wrong ways, how they have been used in this current process?
- Why did two similar cases reach different outcomes? While the first two questions address the inputs effect, the last one is mainly about the process itself.

Explainability, on the other hand, is a property of the autonomous decision-making system, describes how simply an explanation can be generated. Explainability demands for simplicity, which means for a decision-making system, in which a minimum number of assumptions and

input variables that are most relevant and concrete ones have been used, generating a valid explanation would be more straightforward. However, it limits the achievable performance. Hence, Explainability is not cost-free for the designers and, consequently, the users.

Generating explanation not only is not cost-free but also may reveal trade secrets. Therefore, we should know where and when an explanation would be preferable by society or is requested by the law. Followings are some of the situations when the benefits of the explanation dominate the costs:

- Whenever the decision has an impact on another person or any other legal personality.
- If there is value in having an explanation, for example, if one can react based on the law, or one can request for redress.
- The situation when one believes that an error has occurred and having evidence for the inadequacy or unreliability of the input data, and in cases where one is suspicious about the output since, for example, the decision-maker gives a different output for the same cases or the same output for remarkably different cases.
- In cases where one is suspicious about the integrity of a chain of decisions, for example, if the decisions show a remarkable benefit in favor of others.

Nature of the decision, the susceptibility of the decision-maker to outside influence, moral and social norms, the perceived costs and benefits of an explanation, and a degree of the historical accident are also other factors that are addressed by the authors in [33], based on which the society or law may request for an explanation. The agencies should provide at least those explanations requested by the law and are responsible for giving a reasonable explanation when required.

One trend is to equip the autonomous decision-making systems with explanation generators, which are parallel to a central system and provide a reasonable explanation for each outcome or at least to provide some clues regarding how a specific output is related to the input or how a variable will impact the output. Building such a system is again not cost-free, and the main challenge is to build a system capable of generating human-interpretable explanations. In fact, explanations should be such that at least they do not create further challenges in understanding them.

In cases where the financial burden of building explanation generators is not acceptable for the designers, especially for small companies, alternatives are empirical evidence for the system behavior or provide a theoretical guarantee for the system's behavior. However, they cannot be

used in all the cases, as sometimes empirical evidence is not valid, or as in the big AI systems providing a theoretical guarantee is extremely complicated, if not impossible. Nevertheless, being able to generate explanation when and where it is essential is the responsibility of the users, since as discussed before, based on the AIA they should be an expert within the scope of their services.

## AI AND PRIVACY

The previous two sections were mainly related to the regulation of the ongoing impacts of integrating AI solutions with human society, while in the subsequent ones, we will deal with the consequences of design requirements for AI solutions on human society. In this section, the impacts on privacy will be addressed, while in the next section, we will discuss a framework for a better comparison of adopting AI solutions vs. analytic ones. Therefore, these sections mainly focus on the AI developers as the role players.

Data for AI systems is an essential design requirement of AI solutions. Hence, a shift toward AI solutions consequently affects people privacy. The current data protection and privacy laws, which were mainly based on not disclosing the data, are not compatible with these new technologies, and revising and reforming these laws and methodologies are required. Besides, AI solutions may also have ongoing impacts on privacy due to their capabilities in deanonymizing the protected information and mining the complementary data from large data sets [26] , and their computational capacities.

Human interactions with social media, social networks, the general internet, video surveillance cameras, and the general IoT infrastructure have highly increased within the past decade, and it will increase further in future. Besides, dependencies on these tools have also increased in the past few years, which implies that the situation is becoming more stringent for the policymakers and lawmakers for regulating AI solutions. There are two major concerns, first, protecting privacy during data acquisition and in information flows, and secondly to ensure that people have enough control over their data.

The first step is to revisit the conceptualization of privacy, as studied extensively in [8] with the main focus on privacy in the social networks, and Facebook photo tag suggests AI engine as a case study.

As it turned out, the traditional conceptualizations of privacy, i.e., the theory of the right of being left alone, theories based on limited accessibility and secrecy, the right of controlling

personal information, the individuality theory, and the pragmatic approach to privacy are not general enough to cover all the emerging sources of information and has some major deficiencies. Among the remaining, the contextual integrity theory, which has taken into account the users' online privacy more explicitly than others, better fits the context, which also consists of the social networks, internet, and social media.

The contextual integrity theory focuses on the contexts in which the information will be shared and the norms governing those contexts, while both private and public flow of information have been addressed in this framework. The contextual integrity theory goes through the following steps:

- Determining the relevant "context" of the particular flow,
- Identifying the parties involved in the flow, and defining the "sender," the "recipient," and the "subject" of the information,
- Identifying the nature of the information, by considering different information types involved in the flow,
- Identifying the relevant "transmission principles", and
- Determining the applicable "informational norms" that have been developed in an analogous context offline and traditionally in the society and applying the norms to the informational flow.

A flow of information violates privacy by changing the nature of the information, its transmission principles, or making the information available to others who were not involved in the normal flow of information without any owner agreement. The theorem preserves rooms for superior moral or political overruling the flow. At the same time, a change will be considered superior if "it positively affects an important concern for the society, such as "autonomy and freedom, power structures, justice, fairness, equality, social hierarchy, and democracy" in these cases, one should weigh those against the other norms in the context. This conceptualization along with EU GDPR [29], which properly defines the individual and sensitive information and the minimal requirements for accessing and using that information to provide control for individuals over their information can be used for determining privacy breaches by data users.

However, having a covering conceptualization of privacy and data protection are not enough, and as discussed in [27] the approach for protecting privacy should also be revisited. From their work [27], among them we suggest the following to be considered in the privacy frameworks:

**Shifting from individual consent to data stewardship:** In contrary to the current data protection laws, which place most of the responsibility for protecting privacy on the individuals through the consent agreement contracts, the article suggests, in order to be more effective in case of AI systems and enormous data sets available, it would be better to shift more responsibility to data collectors and data users. This, in turn, will also benefit the agencies by not wasting resources on designing and using the AI system restricted by "efforts to comply with ineffective measures, such as notices that no one reads or adhering to terms of consent that are often illusory at best" [27].

**A greater focus on data uses and impacts instead of data collection and analysis:** Under a more usage-based approach, data users would evaluate the intended utilization of personal data not by the terms under which the data has been collected but based on the possible impacts of the current usage and the ongoing risks they pose for individuals. The methodology is superior because it is more relevant to individual's primary goal in saving their privacy by considering any discomfort resulting from processing those data. Therefore, this eliminates the requirement for the purpose specification when collecting the data and shifts the consent agreement to the usage intention instead.

As far as the implementation of regulations are of concern, as discussed in [8], we suggest to jointly use the following tools for forcing the data users to behave appropriately:

- The law, which mandates certain behavior and imposes sanctions on deviate actions,
- Social norms, which impose sanctions when members of a community stray from commonly held expectations of proper behavior,
- The market further tends to inflict a cost on certain actions, incentivizing market participants to modify their behavior, and
- Architecture, which can mandate behavior in the most effective way, by making deviate behavior technically impossible or difficult.

We can use the law to mandate certain behavior and impose sanctions on deviate actions, the market to inflict a cost on specific actions, and incentivize the market participants to modify their behavior. At the same time, one needs to provide the infrastructure that makes the alternative desired architectures possible. Also, by culture building and raising people's awareness about the context, one can use the social norm to force companies to change their behavior in the long term. It should be noted that they are not mutually exclusive, and no single one is sufficient for proper regulation; they should work together to reach the optimal performance.

## ASSESSING THE AI SOLUTIONS

Having an assessment tool helps one to better formulate the debates in using or not using AI solutions. For the particular application of AI in facial recognition systems, such an assessment tool has been proposed in [28], which can be adopted for other applications as well, assuming there is a good understanding of possible harms and benefits. Hence, here we will discuss and present this assessment framework as both an assessment tool for AI applications in facial recognition systems and other problems. Such an assessment can provide a better formulation for comparing AI solutions with analytic solutions for each application.

The framework proposed in [28] will be called hereafter by the abbreviation "FIA", which is an incremental framework for a rigorous and comparative impact assessment of the AI solutions. It is composed of four consecutive steps:

1. **A rigorous study of the declared purpose:** In this step, the declared purpose or the underlying problem for which one seeks the solution will be addressed independent of how it can be solved, no matter if it can be solved with AI or any other analytical solution. For example, the purpose of using AI can be to increase the profit by providing ease of customization or marketing and advertising.
2. **Study of the means to achieve the purpose:** The measurements and control variables associated with the goal, no matter how they can be utilized in the real world, should be addressed in this step. For example, a particular means for advertisement and customization is to track the user's visit to the Amazon website, along with other information, to conclude about his/her attitudes toward specific brands and goods.
3. **Studying the utilization of the AI solutions to reach the goal:** In this step, assuming that the previous steps have already passed, we demand a rigorous risk assessment of using AI solution in the particular context and the possible harms and ongoing risks of using AI, considering the possibility that AI can go beyond what it has been designed for, initially, assuming that the AI solution can reach 100% of its inference performance.
4. **Risk assessment of the implementation of the AI solution:** In this step, the risks associated with different implementations of the AI solution should be taken into account. The risk associated with lack of performance and accuracy, explainability and the likes should be addressed. Risk assessment of alternative solutions instead of AI solutions is a determining parameter in this step, enabling one to adequately understand the benefits and superiority of AI solutions as compared to alternatives if any exists from the realization point of view and based on the maturity of theories behind each of them.

In all the above steps, the related debates should be comparative and rigorous. Both benefits and risks must be assessed in comparison with alternatives. This way, we can decide among the solutions while considering different aspects of solutions and the values of the pros and cons associated with all the solutions in all the steps.

In each step, typically, three types of arguments are possible:

- Arguments already supported by experimental measures or results broadly accepted by the scientific community,
- Arguments that are not validated by sufficient studies but that could be tested, and
- Arguments that are based on subjective or political positions and are not subject to experimental evaluation.

For the analysis to be rigorous, enough evidence and supporting experiments should be provided to ensure that the analysis is not based on positive or negative preconceived ideas, and the study is rigorous and sufficiently concrete.

The ethical matrix can be used during the procedure to manage better the risks and issues that should be considered. The ethical matrix is a table in which rows are associated with the stakeholders, and the columns represent the impacts to be addressed. Three main categories of AI impacts on society, as also discussed in the original article [28], are well-being (health, living conditions, etc.), autonomy (freedom, dignity, etc.), and fairness or justice. However, as suggested in the article, the matrix could go beyond these categories and consider other types of risk, such as the fundamental rights most affected by facial recognition technologies such as privacy, non-discrimination, freedom of expression, and protection of the rights of child and older people.

## CONCLUSION

In this report, criminal liability, accountability and impact assessment, and privacy protection frameworks for further studies in the field of AI public policy have been discussed and presented. In the first section, we provided a general overview of the problems, the influential factors, and the unique characteristics of the AI technologies that make the conventional frameworks for solving similar problems in AI technologies ineffective.

We also provided some examples of failures of the AI technologies that raised alarms for demanding a more careful integration of AI applications with human society, along with some

examples of privacy breaches in AI companies that required revisiting AI technologies and people interactions. Next, the criminal liability of AI systems has been addressed. Instead of applying design failure lawsuits, we discussed and presented a behavioral approach proposed in the literature, which treats AI system in the same way as treating humans when making mistakes, that fits better the context of AI technologies. Yet, we require some further clarifications in the case of direct liabilities and implementation policies.

To ensure a beneficial integration of AI technologies with human society and manage the ongoing risks, AI users and service providers should follow some accountability guidelines. In this report, we mainly focused on the famous algorithm impact assessment framework to be used by AI users as well, while the role of explanation in keeping people and service providers in touch has also been presented.

The criminal liability and accountability frameworks are mainly for managing the ongoing risks of AI solutions. While AI technologies will worsen the privacy issues based on their capability for deanonymizing sensitive information. Firstly, we discussed and presented a proper conceptualization of privacy, which also covers social networks and the internet as the emerging source of information. This conceptualization and proper definitions of individual and sensitive information and the standards required for accessing such information addressed in the famous general data protection and regulatory framework can properly determine the cases of privacy breaches to take care of them.

Simultaneously, to be more AI-friendly, some literature suggestions for adaptation of the privacy frameworks were discussed and presented. Finally, to manage the rush towards AI solutions, a framework previously used for impact assessment of AI-based facial recognition systems has been suggested here for assessing the risks and benefits of applying AI solutions for a particular application. Using such a framework enables one to better understand the advantages and limitations of AI solutions in each application to see whether its application is worth the associated risks as compared to other possible alternative solutions.

## REFERENCES


[1] OECD, Artificial Intelligence in Society, 2019, p. 139.

[2] S. Suwajanakorn, S. M. Seitz and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, p. 1–13, 2017.

[3] Q. Rao and J. Frtunikj, "Deep learning for self-driving cars: chances and challenges," in *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*, 2018.

[4] S. Donnelly, *AI Pilot Successfully Lands Plane Carrying Human Passengers*, 2019.

[5] X. Wu and X. Zhang, "Automated inference on criminality using face images," *arXiv preprint arXiv:1611.04135*, p. 4038–4052, 2016.

[6] L. Johnson and M. Fitzsimmons, *Uber self-driving cars: everything you need to know*, TechRadar, 2018.

[7] J. Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women*, Thomson Reuters, 2018.

[8] Y. Welinder, "A face tells more than a thousand posts: developing face recognition privacy in social networks," *Harv. JL & Tech.*, vol. 26, p. 165, 2012.

[9] A. Marshall, *Uber's Self-Driving Car Didn't Know Pedestrians Could Jaywalk*, Conde Nast, 2019.

[10] J. P. Steil, L. Albright, J. S. Rugh and D. S. Massey, "The social structure of mortgage discrimination," *Housing studies*, vol. 33, p. 759–776, 2018.

[11] N. Confessore, *Cambridge Analytica and Facebook: The Scandal and the Fallout So Far*, The New York Times, 2018.

[12] Y. Bathaee, "The artificial intelligence black box and the failure of intent and causation," *Harv. JL & Tech.*, vol. 31, p. 889, 2017.

[13] J. Clement, *Most used social media 2020*, 2020.

[14] D. A. V. I. D. NIELD, *All the Ways Google Tracks You-And How to Stop It*, Conde Nast, 2019.

[15] J. Lee, *Shopping vs. Privacy: What Does Amazon Know About You?*, 2017.

[16] J. Clement, *Social media - Statistics & Facts*, 2020.

[17] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron and P. Barnes, "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[18] F. Lagioia and G. Sartor, "AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective," *Philosophy & Technology*, p. 1–33, 2019.

[19] G. A. Kaissis, M. R. Makowski, D. Rückert and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, p. 1–7, 2020.

[20] N. A. Greenblatt, "Self-driving cars and the law," *IEEE spectrum*, vol. 53, p. 46–51, 2016.

[21] K. Hao, *AI is sending people to jail-and getting it wrong*, MIT Technology Review, 2020.

[22] D. Coldewey, *AI desperately needs regulation and public accountability, experts say*, TechCrunch, 2018.

[23] J. Millar, B. Barron, K. Hori, R. Finlay, K. Kotsuki and I. Kerr, "Accountability in AI. Promoting Greater Social Trust," in *Conference on Artificial Intelligence, Montreal*, 2018.

[24] A. N. Institute, *Algorithmic Impact Assessments: Toward Accountable Automation in Public Agencies*, Medium, 2018.

[25] D. Li, *It takes a village to protect privacy*, Elsevier, 2019.

[26] B. Lubarsky, "Re-Identification of "Anonymized Data"," *UCLA L. REV*, vol. 1754, 1701.

[27] F. H. Cate and R. Dockery, "Artificial Intelligence and Data Protection: Observations on a Growing Conflict," *Seoul National University, Journal of Law and Economic Regulation*, pp. 1-30, 2018.

[28] C. Castelluccia, "Impact Analysis of Facial Recognition," 2020.

[29] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[30] G. Hallevy, "The criminal liability of artificial intelligence entities-from science fiction to legal social control," *Akron Intell. Prop. J.*, vol. 4, p. 171, 2010.

[31] J. K. C. Kingston, "Artificial intelligence and legal liability," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 2016.

[32] *Strict Liability*, Legal Information Institute.

[33] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger and others, "Accountability of AI under the law: The role of explanation," *arXiv preprint arXiv:1711.01134*, 2017.

[34] D. Reisman, J. Schultz, K. Crawford and M. Whittaker, "Algorithmic impact assessments: A practical framework for public agency accountability," *AI Now Institute*, p. 1–22, 2018.