Machine Learning1 Project Report

# Ethical Issues in Machine Learning

**Dr. Afaf Almehmadi**

Student name: Fatima Ahmed Al-Zahrani ID: 444006628

Student name: Fatima Khaled Al-Amodi ID: 444015415

Student name: Fidaa Osama Flmban ID: 444006581

Student name: Joury Nashat Dome ID: 444007321

Student name: Heba Fahad Al-Matrafi ID: 444010917

## Abstract:

The number of papers being submitted to conferences has sharply increased as a result of recent achievements in the machine learning community. Some of the flaws affecting the present review process employed by these conferences become more visible as a result of this growth. So in this report we focused on the topic of ethical issues in machine learning, and highlighted five vital areas: data collection, human rights and judicial decisions, unfair training practices and bias, responsibility and accountability, and copyright and intellectual property. Our goal is to identify the ethical implications of these topics and their impact on society. One of the primary approaches in this discipline is "Text and Data Mining," which seeks to extract information from texts and digital data through analysis. Systems can use copyrighted data for analysis and learning thanks to this technology. Notable changes include the adoption of exclusions in European law, including the "data mining exception" (TDM) under European directives, which improves the use of protected data for research. But this might result in the usage of private or sensitive data, which would be against the law. Furthermore, depending on skewed or random data might lead to biased conclusions from algorithms, which can be harmful to vulnerable populations.

## 1. Introduction:

The rapid advancement in machine learning and artificial intelligence technologies has revolutionized industries.

It has provided impressive results in data analysis and decision-making, which of course depends on a huge amount of data. Therefore, we have major ethical challenges, concerns about privacy, surveillance, and manipulation of user behavior[3]. The Cambridge Analytica incident was one of the biggest scandals regarding the ethics of machine learning [1].

Artificial intelligence tools were used to influence elections and public opinion, which highlighted the ethical implications of using personal data without consent [6]. This review discusses the ethical issues surrounding machine learning, such as privacy violations, bias, surveillance, and behavioral manipulation, while emphasizing the importance of developing and practicing responsible regulations to protect individual rights [6].

From that aspect it will bring us to another concern of ethical issues related to the application of machine learning in law mainly revolve around concerns such as bias, transparency, accountability, and fairness.

To ensure the ethical use of these technologies, steps must be taken to ensure the models are accurate, fair, and interpretable, while also upholding privacy rights and preventing discrimination. The researchers examine how the online availability of court rulings has enabled the automated analysis of legal data[5]. While automating legal analysis is not a new concept, modern technologies have evolved to include features like automatic summarization and data extraction from legal documents. This review focuses on using natural language processing (NLP) to predict judgments from the European Court of Human Rights [5], highlighting the potential of machine learning in legal contexts. By applying machine learning, the system analyzes the language and terminology used in court cases to learn how to forecast outcomes.

After making predictions, it allows for further examination of the terms that influenced the judgment. It is important to note that this research is based on available datasets and does not claim to predict future human rights rulings [5].

In subsequent sections, the review discusses previous research, explains the use of machine learning for legal text analysis, and presents the experiments and results. Furthermore, these days prediction-based decision algorithms are widely used, especially by governments and organizations.

The main fields where they are applied include lending, contracting, online advertising, and criminal pre-trial proceedings, as well as public health and other areas. With the extensive use and spread of these techniques, ethical issues related to bias in the models and their fairness have emerged [7].

This can diminish the performance of algorithms due to sensitive problems related to race, gender, class, and more. Systems that impact people's lives raise concerns about their ability to make judgments in a fair and unbiased manner [7]. Recognizing and mitigating bias and injustice is quite challenging, as the concept of bias varies from one culture to another.

Additionally, several factors affect the standard of justice, including user experience, culture, and social, legal, and ethical considerations. The review also covers the impact of computer vision and machine learning technologies on many spheres touches upon clear accountability and

responsibility brought up against all stakeholders in the course of technology development and deployment [9]. Otherwise, if system failures and effects could be harmful to individuals and society, then in the absence of such frameworks, public trust may be undermined [9].

The review discusses existing frameworks related to accountability and responsibility relevant to the computer vision and machine learning ecosystem, discusses challenges in the issue of accountability within a multi-stakeholder environment, and calls for proactive approaches toward the governance of mechanisms to ensure that emerging technologies are put to ethical and responsible use [9].

Accountability in AI is an issue because responsibilities are shifted onto the AI systems. All these AI policies talk about the need to develop fair and values-aligned AI systems, as well as appropriate accountability processes [9,10]. Yet, accountability in the context of AI is a very poorly specified term. It does not easily admit public discourses, nor even engagement in policy-making processes.

Such vagueness partly stems from the multifaceted character of accountability, intrinsically complex political processes, and the sociotechnical nature of AIs. The above-said ambiguity in defining accountability in AI, statement of necessary conditions, analysis of architecture, and key accountability goals like compliance and oversight will be attempted to be explained herein. Accountability in AI shall be well understood for proper governance decisions [9,10].

Finally, the review states the "legal responsibility of ethics related to intellectual property" as an important issue in the field of machine learning, as the success of machine learning models heavily relies on the available training data. With the increasing use of data protected by copyright, a serious question arises regarding the permissibility of using this data in the development of machine learning models. Understanding how copyright affects the use of data in machine learning is crucial to ensuring a balance between the rights of creators and the needs of developers. [11][12][14]

## 2. Ethical considerations in machine learning

### 2.1 Data Collection:

The controversy surrounding Cambridge Analytica

The Cambridge Analytica incident is considered to be one of the biggest data privacy breaches in history due to the improper exploitation of private user information for political and commercial gain. 2018 saw the situation come to light after it was discovered that the British political consulting firm Cambridge Analytica (CA) had used the personal information of thousands of Facebook users to affect election campaigns, including the 2016 UK Brexit vote and the US president election[1]. What happened was that

How it Happened:

1. **Data collection**: Cambridge Analytica obtained millions of Facebook users' personal information by using the This Are Your Digital Life app. The program asked users who responded to the quick query for personal information. Unfortunately, the program obtained personal data without permission from approximately 87 million people, not including data taken from users, their acquaintances, and their followers on Facebook[1].

2**. Data analysis**: Using machine learning and data analytics, the company developed comprehensive profiles of each worker based on their behaviour, preferences, political views, and personal characteristics.

3. **Created highly targeted political ads**: Cambridge Analytica ran politically charged ads with the aim of influencing people's opinions and feelings using these profiles. These ads featured politicians or emphasized their positions on important political issues. Subsequent effects:

1. **Impact on Facebook**: Due to the controversy, Facebook came under fire for its role in allowing the hack. Facebook CEO Mark Zuckerberg was questioned by the US Congress about the development of the company's data protection policies.

2. **Political impact:** The controversy fuelled much of the debate about the role of big data in influencing and manipulating public opinion and political outcomes, and highlighted questions about the legitimacy of democratic processes.

3. **Legal and ethical implications**: The incident drew attention to the European Union's General Data Protection Regulation (GDPR) and reinforced the need for stronger data protection rules. The implications for data privacy persisted even after the Cambridge Analytica withdrawal in May 2018. Vigilance, AI, and Privacy The incident demonstrated how easily private information, especially seemingly innocuous information, can be obtained and weaponized for purposes that compromise personal standards and individual freedoms.

Concerns about **privacy and surveillance** have become more important due to the growth of artificial intelligence and information technology, which has led to the collection of vast amounts of data and the creation of rules to protect it Technology: Access to private data and who has access to it is a hot topic when it comes to privacy and surveillance in the digital age. The evolution of technology has outpaced the evolution and increase in laws, leaving a gap for companies, governments and individuals to use data for their own purposes – often without telling anyone[3].

These surveillance issues are compounded by AI-powered capabilities such as voice, facial and fingerprint recognition on devices. The data sets available enable individuals to be monitored, identified and profiled, providing detailed information on personal behaviour – often without much public attention or awareness. Behavioural manipulation: AI has created additional ethical issues and problems in addition to data collection, precisely because of its ability to influence and manipulate behaviour.

AI systems can be used to target, influence and change behaviour both online and offline. This has raised concerns about the weakening of individual autonomy and the emergence of malicious actors who exploit behavioural biases for financial gain. Facebook has become an arena for political manipulation, as the Cambridge case demonstrated.

**Methods**:

**Federated Learning (FL):** This centralized method for learning , such as smartphones or edge devices, to learn models collectively without sending their data to a central repository. Rather, these gadgets do local computations on their input and forward only the modified model parameters to a central server, which merges them to enhance the model as a whole[2] .

Because it guarantees that private user information remains on the device, FL is especially helpful in situations where a high level of privacy protection is required, including in electronic ecosystem and the health care sector.

Self-Supervised

**Learning (SSL):** This machine learning method uses unlabeled data to its advantage by inferring tasks or labels from the data itself. For instance, an SSL model may be trained to recognize [2] images.

**Both approaches address different needs in machine learning (see Fig. 1).**

| Category | Federated Learning (FL) | Self-Supervised Learning (SSL) |
|---|---|---|
| **Advantages** | 1. High Privacy: Data is not removed from the user's device. | 1. Cost: Does not need to name the data; it will cost or take time. |
| | 2. Decentralized Training: No central data needed. | 2. Adaptability: Applies to several types of data (texts, images, etc.). |
| | 3. Reduced Data Transfer: Only model updates are shared, minimizing data leakage. | 3. Feature Learning: Models can learn useful representations from raw data. |
| **Disadvantages** | 1. Complexity in Aggregation: Integrating updates is challenging. | 1. Implementation Challenges: Designing effective self-supervised systems can be complex. |
| | 2. Potential for Inference Attacks: Model updates may inadvertently reveal information. | 2. Risk of Overfitting: Models may overfit to the generated tasks. |
| | 3. Communication Overhead: More clients require more resources to manage updates. | 3. Data Requirements: Needs big data for effective training. |

Figure1: Comparison between FL & SSL

## 2.2 Human Rights and Judicial decisions

The objective of this study is to create a system that applies machine learning techniques to categorize legal documents, specifically focusing on predicting case rulings. The method employed involves supervised learning, where the computer is supplied with text data from court cases and their respective outcomes, allowing the system to detect patterns linked to various types of judgments (e.g., violation or non-violation) [5]. Once the model has been trained, it is evaluated by being presented with new cases without rulings, requiring it to predict outcomes based on the patterns it has acquired during training.

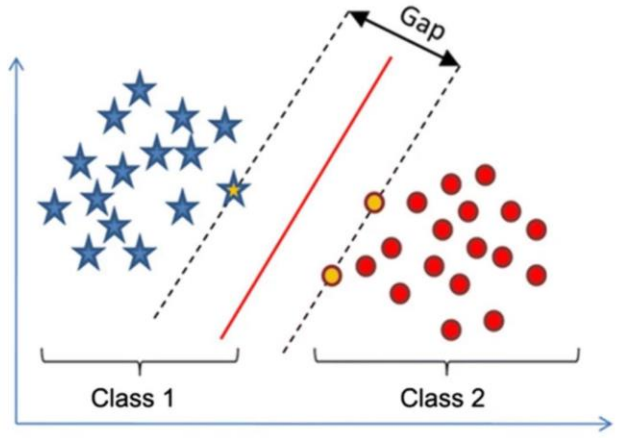


Figure2: SVM dividing data into classes [5]

Once the SVM model is trained, its effectiveness is measured by using a separate test set that was not part of the training data. The system categorizes each case as either a violation or non violation, and these predictions are then compared to the actual court decisions (see Fig. 2). The model performance is the accuracy of its predictions. Another method for evaluating the model's performance is "k-fold cross-validation." [5] Here, the model is trained on k-1 parts after the data is divided into k parts, with the remaining part used for testing. This process is repeated k times, ensuring that each segment is used for both training and testing. For instance, in "five-fold cross-validation," [5](see Fig. 3&4) the model undergoes training and testing five times, with 20% of the data being tested in each round

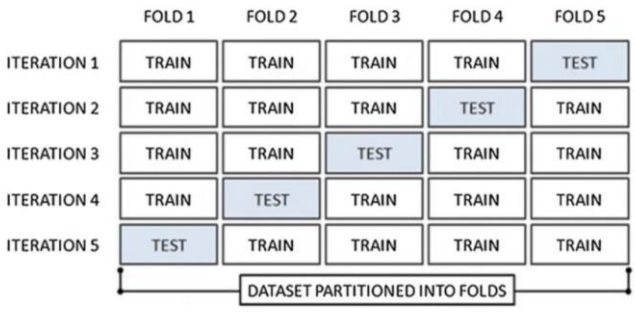


Figure3: Example of fivefold cross-validation [5]

| Advantages | Details |
|---|---|
| **Evaluation Accuracy:** | Separate test set: Enables assessment of the model's performance on unseen data, giving a clearer picture of real-world handling. |
| | K-fold cross-validation: Ensures every portion of the dataset is used for both training and testing, reducing bias from uneven data splits. |
| **Avoiding Overfitting:** | Cross-validation helps prevent the model from learning patterns specific to one dataset, improving generalization to unseen data. |
| **Disadvantages** | |
| **Increased Computational Load:** | Requires retraining the model multiple times (k times), increasing time and computational resources needed for training. |
| **Limited Ability to Generalize:** | Despite using cross-validation or a separate test set, the model might still struggle to predict outcomes accurately if future data differs significantly from the training data. |

Figure 4: Comparison between test sets [5]

## Privacy Attacks in Machine Learning

### **Problem and Definitions**:

Machine learning is a field that focuses on how to teach systems to extract patterns and knowledge from data without having to explicitly program them. The goal here is to provide a simplified overview of machine learning with the aim of preparing the discussion in the following chapters. These are different types of machine learning methods and their classifications, as well as the structures used in these models. In addition, we will briefly discuss the process of training and infering models. For more details, there are several books covering this topic comprehensively.

Types of learning: Traditionally, machine learning is divided into three main areas: guided learning, unguided learning, and enhanced learning [6]. However, in recent years, new types of learning such as semi directed learning and self-learning have emerged, as well as other classifications of models such as generational models and discriminatory models. One of the algorithms

**One of the algorithms:**

The training process usually involves the use of an iterative improvement algorithm such as Gradient Descent (see Fig. 5), which aims to reduce the target function by tracking the path inferred from its gradients. When the volume of data is large, as is usually the case with deep neural networks, taking one step into gradient becomes very expensive. In this case, it is preferable to use modified versions of gradual regression that rely on smaller steps using small batches of data. One of these methods is known as random progressive regression (SGD), where the model is updated using small random batches of data rather than the full data set.

$$\theta_{t+1} = \theta_t - \eta \mathbf{g},$$

$$\mathbf{g} = \frac{1}{m'} \nabla_\theta \sum_{i=1}^{m'} l(f(\mathbf{x}_i; \theta), \mathbf{y}_i),$$

Figure 5: Gradient Descent & step size functions [6]

The most common learning algorithm in federal learning is the "Federal Centering" algorithm (see Fig. 6), where each remote device calculates one step of gradual regression using locally stored data, and then sends the updated values of the model's weights to the central server. The server centers these weights from all remote participants to update the global

model, which is later shared with remote devices again.

$$\theta_{t+1} = \frac{1}{K} \sum_{k=1}^{K} \theta_t^{(k)},$$

Figure 6: Federal Centering function [6]

Attack Types: In privacy-related attacks, the goal of an adversary is to gain knowledge that was not

intended to be shared. Such knowledge can be related to the training data or the model f, or even aspects of the data, such as biases that may have been unintentionally encoded. In our taxonomy, the privacy attacks studied are categorized into four types: membership inference, reconstruction, property inference, and model extraction (see Fig. 7&8).

**Membership Inference Attacks**:

attacks that determine using black-box or white-box approaches whether a specific input is part of the training dataset, affecting both supervised and generative models [6].

**Reconstruction Attacks**:

These attacks aim to recreate training samples or their labels using techniques like model inversion and attribute inference [6].

**Property Inference Attacks**:

These attacks extract dataset properties that are not explicitly encoded, revealing demographic information or other attributes that could have privacy implications [6].

**Model Extraction Attacks**:

These attacks aim to replicate the underlying model by querying it, allowing adversaries to gain insights into its functionality [6].

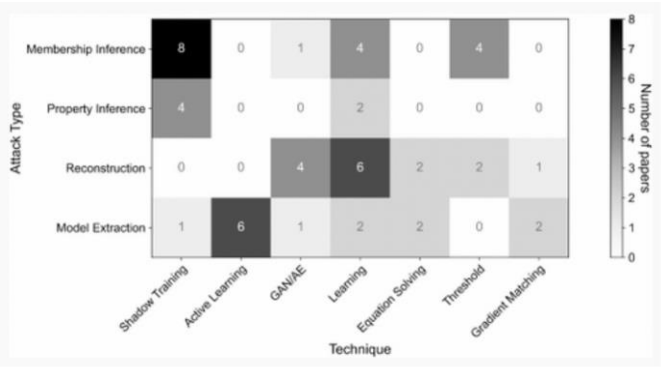


Figure 7: The effect of different techniques on attacks [6]

| Advantages | Disadvantages |
|---|---|
| **Motivation to develop more secure systems**: These attacks can contribute to advancing research and development towards more robust machine learning systems capable of resisting various types of attacks. | **Violation of fundamental rights**: These attacks constitute a clear violation of the right to privacy, as individuals' personal data are exploited without their consent. |
| **Vulnerability detection**: These attacks help reveal hidden security vulnerabilities in machine learning systems, allowing them to be fixed and strengthening protection. | **Systems manipulation**: These attacks can be used to manipulate AI-based decision-making systems, leading to adverse and dangerous consequences. |
| **Stimulating ethical debate**: These attacks spark a lively debate about ethical issues related to the collection and use of data, prompting more stringent ethical standards to protect privacy. | **Destruction of trust**: These attacks reduce users' trust in technology and make them more cautious about sharing their personal data. |

Figure 8: Comparison of the effects of different attacks on ML development

## 2.3 Unfair training and bias

There are three main approaches for designing a fair classifier (see Fig. 9):

| Approach | Description |
|---|---|
| **Pre-processing Approach** | This approach focuses on transforming the training data before it is used. The goal is to improve the data to align with fairness criteria. |
| **In-processing Approach** | This approach modifies the learning algorithm itself to meet fairness criteria. Considerations of fairness are integrated directly into the learning process. |
| **Post-processing Approach** | This approach modifies the decisions made by the classifier after it has been trained. This means that the final outcomes are altered to ensure fairness. |

Figure 9: approaches for designing a fair classifier

The importance of minimizing the risks of erroneous or biased decisions is increasing daily, particularly in sensitive areas, due to our growing reliance on these systems. As a result, obtaining a transparent algorithm that can explain how and why it made its decisions fairly and without bias has become increasingly challenging.

Moreover, models must meet specific requirements for interpretation, such as local interpretation (focusing on the important factors in a particular decision) [8] and global interpretation (evaluating all decisions based on certain criteria) [8].

Additionally, a mental model can be employed to assess whether the model can accurately predict classifications. If the model can predict the classifications of the main model, it is on the correct path to transparency. [8] It is important to mention that machine learning models can be classified as either "white-box" or "black-box," depending on their availability and constraints.

White-box models are machine learning models that facilitate understanding due to their accuracy and transparency. They also offer flexibility in structural adjustments, making it easier to modify and change the model because of their simplicity [8].

Black-box models are machine learning models that are difficult to interpret and understand, and they cannot explain the reasons behind their decisions, resulting in lower transparency. Additionally, there are limitations on the ability to change and improve their structure due to their complexity [8].

## In terms of techniques

There is the category of distributionally robust optimization (DRO), where the goal is to minimize the worst-case training loss[8]. This means ensuring that the model does not incur significant losses even in the presence of changes in data or conditions.

Various types of metrics have been considered to determine the proximity between distributions. Among these metrics are bounded f-divergence and Wasserstein distance [7], which are important in many applications.

Now, how can new techniques be used to discover discrimination in systems?

## Problem and Definitions:

In the context of designing classifiers to address the unfairness gap, various fairness metrics are defined along with methods to achieve them through specific mathematical approaches.

Weight w

δ_w^DP(h) (see Fig. 10), measures the maximum weighted difference in acceptance rates between the two groups with respect to the distribution that assigns weight w to the training examples.

δ_w^EO(h) (see Fig. 11). represents the unfairness gap under the fairness constraint F.

$$\delta^w_{DP}(h) = \max_{a,a' \in \mathcal{A}} \left| \frac{\sum_{i:a_i=a} w_i h(x_i, a)}{\sum_{i:a_i=a} w_i} - \frac{\sum_{i:a_i=a'} w_i h(x_i, a')}{\sum_{i:a_i=a'} w_i} \right|.$$

Figure 10: the function of maximum weighted difference [7]

where δ_w^EO(h|y) is defined as

$$\delta^w_{EO}(h|y) = \max_{a,a' \in \mathcal{A}} \left| \frac{\sum_{i:a_i=a,y_i=y} w_i h(x_i, a)}{\sum_{i:a_i=a,y_i=y} w_i} - \frac{\sum_{i:a_i=a',y_i=y} w_i h(x_i, a')}{\sum_{i:a_i=a',y_i=y} w_i} \right|.$$

Figure 11: unfairness gap function [7]

δ_w^EO(h|0) (resp., δ_w^EO(h|1)) measures the weighted difference in false (resp., true) positive rates between the two groups with respect to the weight w.

The Main Objective:

Our goal is to solve the following min-max problem (see Fig. 12).

$$\min_{h \in \mathcal{H}_{\mathcal{W}}} \max_{w \in \mathcal{W}} \ell(h, w)$$

Figure 12: min-max problem [7]

aiming to minimize a robust loss with respect to a class of distributions indexed by W. Additionally, we also aim to find a classifier that is fair with respect to such perturbations.

Design

we demonstrate how to create a fair classifier that excels in accuracy for a specific weight vector w ∈ W, while also ensuring fairness across the broader set of weights W (see Fig. 13&15).

| Component | Meta Algorithm | Approximate Fair Classifier | Setting up a Two-Player Zero-Sum Game |
|---|---|---|---|
| **Objective** | Simplify a complex optimization problem | Balance accuracy and fairness | Formulate a strategic interaction between players |
| **Methodology** | Transform min-max to loss minimization | Discretize weights and use Lagrangian multipliers | Create a zero-sum game framework |
| **Complexity** | Relies on approximate Bayesian oracle | Requires cost-sensitive classification | Involves cooperative strategies |
| **Flexibility** | Applicable to various complex problems | Ensures fairness across weight sets | Depends on player interactions |

Figure 13: Comparison of different algorithms for solving min-max problem [7]

**The meta-algorithm** is the main algorithm that Focuses on simplifying and facilitating the min-max problem (see Fig. 14).

**ALGORITHM 1:** Meta-Algorithm

**Input:** Training Set: $\{x_i, a_i, y_i\}_{i=1}^n$, set of weights: $\mathcal{W}$, hypothesis class $\mathcal{H}$, parameters $T$ and $\eta$.

Set $\eta = \sqrt{2/T_m}$ and $w_0(i) = 1/n$ for all $i \in [n]$

$h_0 = \text{ApxFair}(w_0)$ /* Approximate solution of $\arg\min_{h \in \mathcal{H}_\mathcal{W}} \sum_{i=1}^n \ell(h(x_i, a_i), y_i)$. */

**for** *each* $t \in [T_m]$ **do**

  $w_t = w_{t-1} + \eta \nabla_w \ell(h_{t-1}, w_{t-1})$

  $w_t = \Pi_\mathcal{W}(w_t)$ /* Project $w_t$ onto the set of weights $\mathcal{W}$. */

  $h_t = \text{ApxFair}(w_t)$ /* Approximate solution of $\min_{h \in \mathcal{H}_\mathcal{W}} \sum_{i=1}^n w_t(i) \ell(h(x_i, a_i), y_i)$. */

**end**

**Output:** $h_f$: Uniform distribution over $\{h_0, h_1, \ldots, h_T\}$.

Figure 14: meta-algorithm [7]

### Approximate Fair Classifier:

The Approximate Fair Classifier is designed to sort different weights and uses Lagrangian multipliers to determine fairness factors and conditions [7].

### Setting up a Two-Player Zero-Sum Game:

In this technique, the problem of designing a fair classifier is formulated as a two-player zero-sum game between the learner and the adversary .

**Learner**: This player is focused on selecting a hypothesis (a model or decision rule) that minimizes the risk associated with classification errors.

**Adversary**: The adversary's role is to challenge the learner by identifying the weight or scenario that leads to the greatest unfairness in the classifier's decisions [7]

| Component | Benefits | Drawbacks |
|---|---|---|
| **Meta Algorithm** | - Simplifies complex optimization problems | - May not guarantee fairness across all weights |
| | - Flexible application to various classification tasks | - Relies on the availability of an accurate oracle |
| **Approximate Fair Classifier** | - Balances accuracy and fairness | - May require additional computational resources |
| | - Provides a clear mechanism for fairness | - Complexity in designing the classifier |
| **Setting up a Two-Player Zero-Sum Game** | - Strong theoretical framework for interaction | - Can be complex and challenging to implement |
| | - Encourages strategic interactions between players | - May require careful tuning of parameters |

Figure 15: advantages-dis advantages Comparison of different algorithms for solving min-max problem [7]

## 2.4 Copyright and Intellectual Property

### Current State of the Art

In recent years, the topic of copyright and ethics related to intellectual property in the context of machine learning has garnered significant attention. The main methodologies employed in this field include "Text and Data Mining," which is texts and information through technical analysis for the purpose of requesting information. This allows systems to exploit copyrighted data for learning and analysis purposes. Among the notable developments, in Europe, exceptions have been introduced in legislation such as the "data mining exception" (TDM), which enhances the ability to use protected data in research. This can lead to the use of personal information or sensitive data of individuals. This can cause a violation of privacy. Additionally, using random or distorted data can cause algorithms to yield biased results, negatively impacting vulnerable groups. [11]

However, legal and ethical challenges remain. For example, the Fair Use doctrine allows the use of copyrighted materials without the need for the owner's consent, based on the idea that the copy serves a different function than the original work and does not act as a substitute for it, which is also known as transformational use. This law has impacted developments in artificial intelligence regarding how intellectual property rights are understood and could potentially violate the rights of creators. Conversely, this principle can be exploited by large corporations, leading to the erosion of the cultural value of original works, affecting cultural diversity and creativity, and also reducing opportunities for small companies and creative individuals to receive fair compensation for their work, potentially diminishing their scientific incentives.[11]

In general, the importance of copyright emerges in the context of protecting public interests, as it contributes to promoting creativity and innovation by ensuring the rights of creators. These rights are not only a means of protecting individual interests, but also play a vital role in promoting culture and knowledge in society. The responsibilities associated with copyright require a balance between protecting individual rights and the needs of society as a whole. This balance ensures that there is room for expression and creativity while preserving the rights of creators. Therefore, it is imperative for governments and stakeholders to update legal frameworks to ensure effective protection of copyright, thus ensuring continued innovation and cultural diversity in the digital age.[12]

### Different approaches:

In examining the different strategies for tackling copyright challenges in machine learning, two primary approaches emerge. The first approach emphasizes "facilitating access" to copyrighted data, aiming to eliminate legal obstacles that hinder the use of such data. This facilitates easier access for developers and researchers to the vast datasets required for training machine learning models. The second approach is centered on "protecting copyright" (see Fig. 16).

| Aspect | Advantages | Disadvantages |
|---|---|---|
| **Facilitating Access** | - Fosters innovation and research | - May lead to copyright violations |
| | - Facilitates access to large data needed by machine learning systems | - Can create legal uncertainty about acceptable usage |
| | - Supports startups and small developers in creating new solutions | - Allows the use of "low-quality" data, potentially affecting model performance |
| | - Can lead to improved quality of data used in models | - Can lead to unfair exploitation of protected data |
| **Protecting Rights** | - Protects creators' rights and ensures compensation for their work | - Can hinder technological progress and increase development costs |
| | - Maintains creative integrity and encourages innovation by protecting ideas | - Imposes restrictions on data access, which may slow down innovation |
| | - Provides a clear legal framework for developers | - May reduce the diversity of data available for use |
| | | - Requires significant resources to secure legal rights |

Figure 16: Compare between facilitating access and protecting rights [11]

## 2.5Accountability

In this section you can include the following:

**Accountability as Answerability**

Defined as the obligation to inform about and justify one's conduct to an authority **[9]**.

Involves three necessary conditions: authority recognition, interrogation, and limitation of power **[9]**.

**Accountability Architecture**

In order to comprehend what the answerability includes one must take into account:

**1.Accountability Architecture [9]**

Context: The area or context in which accountability is established.

Range: The specific tasks, actions, or decisions for which an agent is accountable.

Agent: An active entity, either individual or group performs actions.

Forum: The entity to whom the agent is accountable.

Standards: The norm or principles used as a point of reference against which the agent's acting can be assessed.

Process: The procedures through which accountability is enforced.

Implications: The consequences of accountability assessment results.

**2. Sociotechnical Approach [9]**

Range: Activities that range from design, development to deployment of the AI life cycle.

Agents: Those who design, develop, and deploy a particular AI system.

Standards: Legal, Ethical, Technical Requirements.

Procedure: Internal audits, external audits, and human-machine interfaces.

Implications: Recommendations, approvals, refusals, and sanctions.

**Goals of Accountability**

Four cardinal goals that shape how accountability is framed by those in governance include compliance, reporting, oversight, and enforcement. These are considered to drive the formulation of accountability regimes, which are presented below in roughly logical order.

**Compliance & Report**

Compliance aims to ensure alignment with ethical and legal standards, while the report goal involves documenting and justifying the agent's actions for review by the forum or principal, aiding in challenging misconduct based on relevant information **[9]**.

**Oversight & Enforcement**

Oversight involves examining evidence and evaluating conduct, allowing for scrutiny of decision-making processes, while enforcement determines consequences based on gathered evidence **[9]**.

**Policy-Maker Considerations**

Policy-makers can pursue these accountability goals individually or simultaneously, with an emphasis on different goals based on legislation and governance factors. This approach is important for effective governance, especially in areas like AI **[9].**

**Descriptive and Normative Rationales**

descriptive and normative reasons for prioritizing certain accountability goals over others, emphasizing the significance of a goal-based analysis for effective governance and political coordination **[9].**

**Comparative Analysis:**

**Advantages**

The literature demonstrated three advantages of the accountability in the life cycle of machine learning systems:

**1-Define some of the major stakeholders and their roles throughout a machine learning system's life cycle [9] (see Fig. 17).**

| Stakeholder | Roles |
|---|---|
| **Developers** | Responsible for the design, training, and implementation of models. They should carry out all tasks with more precision, randomness, transparency, elimination of biases. |
| **Deployers** | Integrate systems into products or services, ensuring proper use, monitoring, and maintenance in line with ethical principles and legal requirements. |
| **Users** | Interact with and rely on the systems, using them as intended and reporting issues while being aware of limitations and biases. |
| **Regulators** | Establish guidelines, standards, and regulations, overseeing development and use to protect public interests and ensure compliance. |

Figure 17: stakeholders and their roles throughout a machine learning system's life cycle **[9]**

**1-Define the policies and procedures [9]**

**Proactive Accountability**
Focuses on preventing or reducing issues.
**Ethical Guidelines and Principles**

Articulate fundamental values such as fairness, transparency, privacy, and human-centeredness to guide development and deployment.

**Standards and Best Practices**

Define requirements for development, testing, and deployment, promoting consistency, interoperability, and quality assurance.

**Regulations and Policy Frameworks**

Establish legal requirements, define liability and redress mechanisms, and provide enforcement powers to regulatory bodies.

**1-Promoting Cooperation and share Responsibility [9]**

**Multi-Stakeholder Engagement**

Encourage ongoing dialogue and collaborative initiatives among developers, deployers, users, regulators, and civil society.

**Education and Awareness**

Promote public education about the capabilities, limitations, and impacts of these systems to foster informed discourse and trust.

**Continuous Monitoring and Improvement**

Implement processes for regular assessment and iterative refinement of systems and governance frameworks.

**Disadvantages**

On the other hand the disadvantages during the life cycle of machine learning systems:

**Complexity of Systems [9]**

The opacity and lack of interpretability of many models, especially deep learning algorithms, make it difficult to trace decision-making processes and identify sources of errors or biases.

**Multi-Stakeholder Involvement [9]**

The involvement of multiple stakeholders can lead to a diffusion of responsibility, complicating the determination of accountability.

**Unintended Consequences and Emergent Behaviors [9]**

Systems can exhibit unexpected behaviors, making it challenging to assign responsibility for unforeseen outcomes.

**Harm to Individuals and Society [9]**

Biased or discriminatory outputs can lead to wrongful arrests, misdiagnosis, and perpetuation of systemic inequalities.

**Erosion of Public Trust [9]**

Repeated biases, errors, or unethical behaviors can diminish public confidence in these technologies.

**Legal and Financial Liabilities [9]**

Stakeholders may face lawsuits, regulatory sanctions, or financial penalties in the absence of clear accountability frameworks.

## 3. Conclusion:

The need for ethics in machine learning is growing exponentially. We need ethical guidance, especially as concerns such as privacy violations, racist algorithms, and manipulation have become widespread. While machine learning is important for predicting outcomes and events, it is also vulnerable to attacks. Addressing the downsides and issues such as bias, unfairness, and transparency is crucial for ethical use.

Accountability and responsibility are essential in computer vision and machine learning. We need proactive governance to define roles for stakeholders, regulators, developers, and users. Intellectual property rights also pose challenges because some have access to copyrighted data. To do justice to creators and developers, it is necessary to balance understanding copyright laws.

I suggest that future research will focus on examining the impact of unfair data use on society, the limitations of low-quality data, and legal loopholes that hinder innovation. Addressing these issues will

ensure the ethical and responsible development of machine learning.

## 4. References:


[1] (N.d.). Tudelft.Nl. Retrieved October 13, 2024, from https://research.tudelft.nl/en/publications/machine-learning-ethics-and-law

[2] Talaei Khoei, T., Ould Slimane, H., & Kaabouch, N. (2023). Deep learning: systematic review, models, challenges, and research directions. *Neural Computing & Applications*, *35*(31), 23103–23124. https://doi.org/10.1007/s00521-023-08957-4

[3] (N.d.-b). Stanford.edu. Retrieved October 13, 2024, from https://plato.stanford.edu/entrieS/ethics-ai/#ManiBeha

[4] (N.d.-c). Datamation.com. Retrieved October 13, 2024, from https://www.datamation.com/big-data/data-collection-trends/

[5] Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, *28*(2), 237–266. https://doi.org/10.1007/s10506-019-09255-y

[6] Rigaki, M., & Garcia, S. (2024). A survey of privacy attacks in machine learning. *ACM Computing Surveys*, *56*(4), 1–34. https://doi.org/10.1145/3624010

[7] Mandal, D., Deng, S., Jana, S., Wing, J., & Hsu, D. J. (2020). Ensuring fairness beyond the training data. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 18445–18456). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/d6539d3b57159babf6a72e106beb45bd-Abstract.html

[8] Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Cruz, G. O. R., Peixoto, R. M., de Sousa Guimarães, G. A., dos Santos, L. L., Araujo, M. M., Marco, C., de Oliveira, E. L. S., Ingrid, W., & Nascimento, E. G. S. (2022). Bias and unfairness in machine learning models: a systematic literature review. In *arXiv [cs.LG]*. http://arxiv.org/abs/2202.08176

[9] V. A. Tuan, "Frameworks for accountability and responsibility among stakeholders in Computer Vision Machine Learning Development and deployment," International Journal of Machine Intelligence for Smart Applications, https://dljournals.com/index.php/IJMISA/article/view/5 (accessed Oct. 13, 2024)

[10] C. Novelli, M. Taddeo, and L. Floridi, "Accountability in artificial intelligence: What it is and how it works," *AI & SOCIETY*, vol. 39, no. 4, pp. 1871–1882, Feb. 2023. doi:10.1007/s00146-023-01635-y

[11] (N.d.-e). https://doi.org/10.12681/bioeth.39041

[12] Czetwertyński, S. (2017). Importance of copyrights in online society. *Managerial Economics*, *18*(2), 147. https://doi.org/10.7494/manage.2017.18.2.147

[13] (N.d.-d). Quillbot.com. Retrieved October 13, 2024, from https://quillbot.com/paraphrasing-tool

[14] *Poe - fast, helpful AI chat*. (n.d.). Poe.com. Retrieved October 13, 2024, from https://poe.com/

## Task distribution:

| Student Name | Task | | |
|---|---|---|---|
| Fatima Al-Zhrani | Unfair training and bias | Main body | Organize the report |
| Fatima Al-Amodi | Data Collection | Conclusion | |
| Heba Al-Matrafi | HumanRightsand Judicial decisions | Security of data | |
| Joury Dome | Responsibility and Accountability | Introduction | |
| Fidaa Flmban | Copyright and Intellectual Property | Abstract | Organize the report |