Addressing the Ethical Implications of AI Models Developed: A Case Study of Master's Degree Dissertations in Data Science for Industry and Society

Alina Delia Călin[a]

Faculty of Mathematics and Computer Science, Babeș-Bolyai University, Kogălniceanu Street, Cluj-Napoca, Romania

Keywords: Ethical AI, Data Science, Responsible AI, Bias, Transparency, Accountability, Fairness, Explainable AI, Safety, Data Privacy, Machine Learning, Artificial Intelligence.

Abstract: The increase in the development and use of AI models has generated many ethical and societal concerns. In this paper, we examine the ethical element in several dissertations presented in July and September 2024 by students enroled in the Data Science for Industry and Society Master's Degree Programme. We assess the level of awareness of ethical principles by analysing in these case studies the ethical concerns addressed by most students, the ethical principles that are mostly neglected, and possible implications for the society. The findings reveal that data bias is the most addressed concern, while accountability is the most neglected ethical principle. Some recommendations for possible improvements include the use of ethical AI tools for the design and assessment of AI models and applications.

1 INTRODUCTION

The past decade has known an exponential growth of Artificial Intelligence (AI) based applications. Many of these, for example public Large Language Models (LLM), introduce disruptive technologies that challenge and modify many aspects of life. The activities involved include teaching and learning, producing written content such as essays, reports, or news, and constitute a huge knowledge and decision base for many users. Despite many disclaimers from the manufacturers of these tools explaining that erroneous content might be generated, many users are trustingly using them in professional and learning contexts. Despite being scrutinised, the use of these models in education to produce writing that is creative, cohesive, and pertinent is deeply explored, emphasising their role in improving the quality and efficiency of student work (Kenwright, 2024).

More often than not, for many such applications, the level of accountability demonstrated by the developer is often neglected, with the result that the user has the primary responsibility to use the AI (Vakkuri, 2022). Caution and reluctance are often advised because the ethical and safety implications are not made known to the user, who is at high risk. Rather than spending too much controversial time on whether we should use and trust these systems or not we might focus more productively on developing ethical models. Promoting transparency, agreeing on safety and inclusive guidelines to be enforced, could finally make this technology serve actual societal needs and improve life quality (Remian, 2019).

The correct identification of ethical concerns is the first step in addressing them. For example, some researchers apparently argue that unemployment as a result of the automation power of AI models is one of the main ethical implications (Wiesenthal, 2022). This calls for a shift in mindset and culture, to employ trustworthy mitigation solutions. Even if these issues are identified and discussed, few studies offer recommendations on how to deal with them when designing and developing AI systems (Huriye, 2023). Moreover, there isn't always agreement among scholars as to what constitutes best practices for ethical AI models (Jobin et al., 2019).

The quality of AI is determined by the availability and quality of data, and it is where the majority of problems originate. It is critically important to ensure that AI systems are trained on a variety of representative datasets and that bias is rigorously tested in them (Konidena et al., 2024). Moreover, it has been demonstrated that AI attacks that utilise gradient-based enhancements are possible (Rosenblatt et al., 2023) resulting in falsely highly accurate AI models.

Real-world case studies offer valuable insight into the successful integration of ethical considerations in

[a] https://orcid.org/0000-0001-7363-4934

Călin, A. D.
Addressing the Ethical Implications of AI Models Developed: A Case Study of Master's Degree Dissertations in Data Science for Industry and Society.
DOI: 10.5220/0013112000003890
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 1*, pages 279-286
ISBN: 978-989-758-737-5; ISSN: 2184-433X
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

AI and Data Science (Tatineni, 2019). Examples include proactive bias mitigation, transparent decision making, and community participation. These strategies ensure fair outcomes for diverse user groups, foster trust, and align ethical considerations with societal values, thereby fostering a more inclusive and ethical AI landscape. Some researchers study how engineering requirements can help businesses handle ethical concerns with AI technologies (Balasubramaniam, 2019), by examining three Finnish firms' AI ethics guidelines using a multiple-case research methodology. Accountability, justice, privacy, safety, security, transparency, and trust were the main topics of the recommendations. Specific suggestions are creating multidisciplinary development teams to address divergent ethical perspectives and prioritising quality requirements by using AI ethical norms.

The current literature has recognised the following major categories of ethical issues: privacy, autonomy, anonymity, transparency, security, safety, justice, and dignity (Tatineni, 2019). These should normally come on top of software engineering quality requirements of usability, performance, reliability, security, safety, maintainability, accuracy, interoperability, and reusability (Balasubramaniam, 2019).

The ethical considerations in AI and Data Science development are crucial for responsible use. Strategies include conducting ethical impact assessments, developing agile ethical frameworks, and fostering collaboration between researchers, ethicists, policymakers, and industry experts. By understanding the implications of advanced AI, anticipating future challenges, and implementing proactive measures, we can contribute to responsible and beneficial technology development. Ethical considerations in AI and Data Science development are crucial for responsible use, requiring impact assessments, agile frameworks, and collaboration among researchers, ethicists, policymakers, and industry experts.

The aim of this paper is to analyse the existing level of knowledge and awareness of ethical AI implications among master's degree students from a specialised Data Science domain. Several thesis are being assessed on 4 selected ethical criteria: (1) fairness and bias, (2) safety and security, (3) accountability and liability, and (4) transparency and explainability. The analysis will focus on the following research questions (RQ): RQ1: *Which ethical principles for developing AI-based applications are most often addressed by students?*; RQ2: *Which ethical principles for developing AI-based applications are generally neglected by students?*; RQ3: *What solutions for the implementation of ethical AI principles from the literature could be applied in these case studies?*.

2 MATERIALS AND METHODS

2.1 Methodology

In this study, we evaluate the ethical element in several dissertation theses presented in July and September 2024 by students enroled in the Data Science for Industry and Society Master's Degree Programme of the Faculty of Mathematics and Computer Science from Babeș-Bolyai University (UBB) in Cluj-Napoca. The criteria used in the analysis of these case studies are: (1) fairness and bias, (2) safety and security, (3) accountability and liability, and (4) transparency and explainability. All identified concerns are then assessed as to how they are addressed in the master thesis, if at all, by means of: discussion, evaluation tools utilised, ethical design, and other approaches.

From a number of 12 student papers, we selected those that involve either the development of an AI module or the development of an application that uses pre-trained AI models. Papers that do not have as scope a specific application addressed to a large pool or users, but are focused more on research methods, comparisons, or assessment of AI models have been excluded. The focus of this study is to assess the level of awareness and mitigation of ethical issues related to developing consumer-oriented applications, and their impact in society.

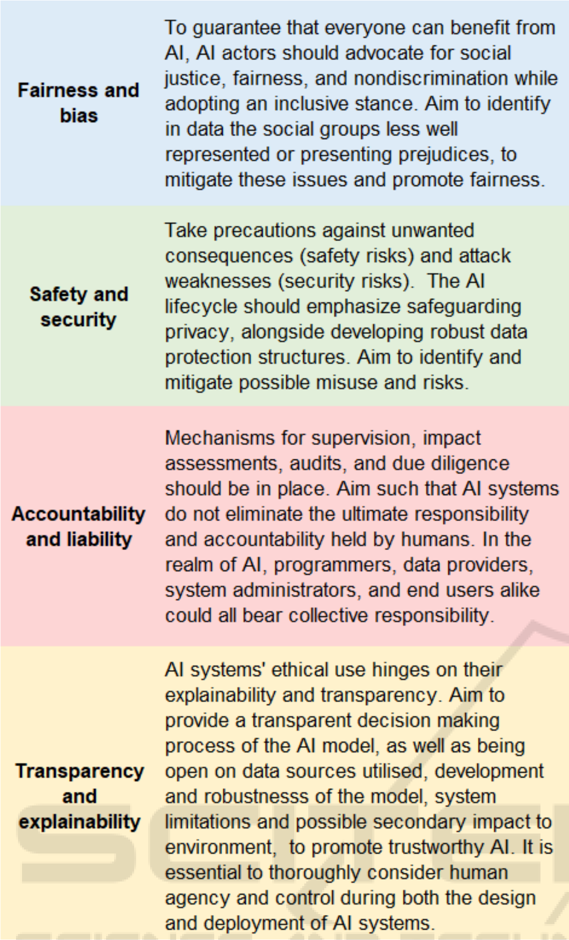
The selected dissertations are being assessed considering the four criteria that are derived from both the related literature, and from ethical AI norms and recommendations, by international organisations . Several concerns specific to each AI model or application are identified and evaluated if they are addressed in the paper, in order to examine the ethical AI literacy among these students and propose several recommendations based on literature.

2.2 Ethical AI Principles

The criteria chosen for this study are based on related literature (Horváth, 2022), (Konidena et al., 2024), (Nguyen et al., 2023) and in line with the recommendations provided by UNESCO's "Ethics of Artificial Intelligence" (UNESCO, 2024). The 4 chosen criteria are described in Figure 1, reflecting what we are focusing on in the analysis of the thesis and identification of possible concerns.

2.3 Summaries of the Selected Papers

Case Study 1 (CS1): Speed Bump Detection. (Complete Title: Robust Speed Bump Detection



Fairness and bias	To guarantee that everyone can benefit from AI, AI actors should advocate for social justice, fairness, and nondiscrimination while adopting an inclusive stance. Aim to identify in data the social groups less well represented or presenting prejudices, to mitigate these issues and promote fairness.
Safety and security	Take precautions against unwanted consequences (safety risks) and attack weaknesses (security risks). The AI lifecycle should emphasize safeguarding privacy, alongside developing robust data protection structures. Aim to identify and mitigate possible misuse and risks.
Accountability and liability	Mechanisms for supervision, impact assessments, audits, and due diligence should be in place. Aim such that AI systems do not eliminate the ultimate responsibility and accountability held by humans. In the realm of AI, programmers, data providers, system administrators, and end users alike could all bear collective responsibility.
Transparency and explainability	AI systems' ethical use hinges on their explainability and transparency. Aim to provide a transparent decision making process of the AI model, as well as being open on data sources utilised, development and robustnesss of the model, system limitations and possible secondary impact to environment, to promote trustworthy AI. It is essential to thoroughly consider human agency and control during both the design and deployment of AI systems.

Figure 1: The Ethical AI criteria used for assessment.

Based on Data Collected from GPS-Enabled Dashcams) (Muntean, 2024). Using predetermined GPS locations, this dissertation proposes a novel way for Advanced Driver Assistance Systems (ADAS) to gather speed bump data. A labelled dataset is constructed with 2351 photos of plastic speed bumps that are black and yellow, used to train a speed bump detection AI model. The research shows that, while maintaining a lower False Positive Rate (of 3.4%), YOLOv8, a cutting-edge object identification model, can achieve performance levels comparable to the state of the art.

Case Study 2 (CS2): Dental Issues Identification. (Complete Title: Advancements in Dental Care through Deep Learning) (Moldovan, 2024). In order to detect frequent dental problems (cavities, implants, impacted teeth, and fillings), this study examines pre-trained ResNet models and custom models trained with TensorFlow Keras. YOLOv9 and other YOLO models are tested for real-time processing on mobile devices. The study showcases the superior performance of the YOLOv9 model in dental X-ray processing, demonstrating its potential to expedite patient sessions and enhance oral health insights with a mobile application.

Case Study 3 (CS3): News Analysis. (Complete Title: Beyond the Headlines: Leveraging AI to Streamline News Analysis) (Miclea, 2024). The project's goal is to create an intelligent news aggregation and analysis system that uses artificial intelligence and natural language processing to automate the task. The system consists of a sentiment analysis module, data processing engine, news retrieval module, and user interface.

Case Study 4 (CS4): Anime Recommendation. (Complete Title: Anime recommendation system) (Lăzărescu, 2024). This thesis compares the results of experiments conducted on a small real-world dataset with the complete MyAnimeList database for the purpose of designing an application to generate anime recommendation. Three models were evaluated: NCF,Wide & Deep, and SVD, with the best result for $R^2 = 0.5473$. The authors' research is a basis for creating an app for anime recommendations.

Case Study 5 (CS5): Stock Trend Prediction. (Complete Title: Predicting Romanian stock movement trends: a complex network approach combined with machine learning) (Holgyes, 2024). This study concentrates on predicting the Bucharest Stock Exchange (BSE) movements. It examines stock price data using network science principles, deriving specific features from the network model (centrality measures and connectivity properties), in order to train several regression algorithms. The top performance was achieved using decision tree, which predicts the following day's price movements (rise, fall, or stagnate) with accuracy of 70% and precision of 50%.

Case Study 6 (CS6): Blood Cell Classification. (Complete Title: Blood cells classification for Acute Lymphoblastic Leukemia detection using federated learning) (Chiorean, 2024). This research investigates the application of federated learning for classifying white blood cells to detect Acute Lymphoblastic Leukemia (ALL). The paper uses Convolutional Neural Networks (CNNs) and the advancements provided by MobileNet and ResNet architectures. The model was able to improve performance up to 90% for accuracy and 80% for precision. ResNet demonstrates superior precision and reliability in diagnosing ALL, while MobileNet excels in performance with a smaller number of parameters.

Case Study 7 (CS7): Travel Recommendation. (Complete Title: A Travel Recommendation System Based on Weather Data and Traveller Profile Using Machine Learning Algorithms) (Zbârcea, 2024). This dissertation aims to assist users in choosing the ideal vacation spot tailored to their individual preferences and requirements. It introduces a web-based platform that provides customised recommendations, valuable insights, and weather-related functionalities. The system employs Long Short-Term Memory models for predicting weather conditions in conjunction with a deep learning model trained on a dataset of virtual users with preferences and past travel experiences.

3 RESULTS AND DISCUSSION

A total of **83** ethical AI concerns have been identified from these papers. One paper of the seven, namely CS1, is distinguished as having addressed most of the safety and security and some of the fairness and bias related issues (9 issues out of 21), while the others fail to address ethical concerns in most aspects, as described in Figures 2 and 3. Also, CS1 has the most ethical concerns addressed in total perhaps because the application domain (automotive) is a popular field studied worldwide. The percentage of issues identified per case study and criteria are presented in Figures 4 and 5, showing that the majority of concerns are related to fairness and bias, and the fewest are related to accountability and liability. This is perhaps because data are the first source of bias in a model. Although accountability is not less important, it is a more straightforward issue of responsibility for the technology and its use.

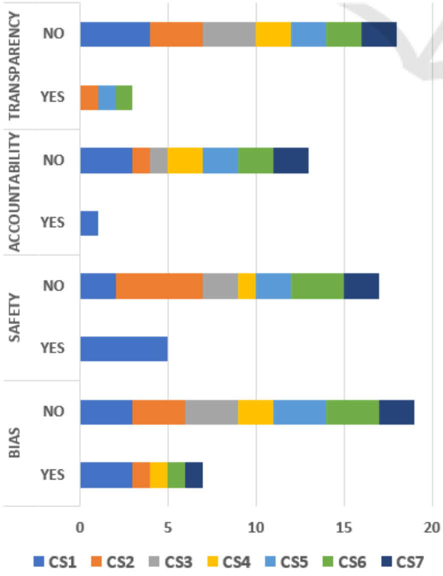


Figure 2: Number of addressed concerns per criteria for each case study.

In Table 1 we present in detail several concerns related to fairness and bias, as they were identified in each of the 7 dissertations, and if they are addressed in any way (either just identified, perhaps discussed, or even mitigated) by the student. A similar analysis for AI safety and security is performed in Table 2. Table 3 presents the issues related to accountability and liability of AI systems, while Table 4 is concerned with transparency and explainability of the models.

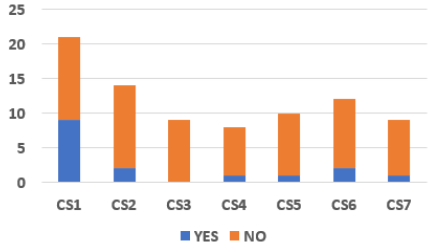


Figure 3: Number of concerns for each case study (yes - addressed, no -not addressed).

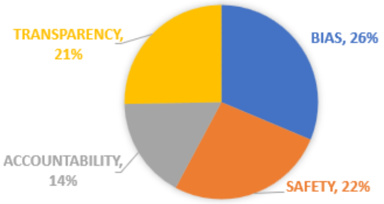


Figure 4: Distribution of concerns per criteria .

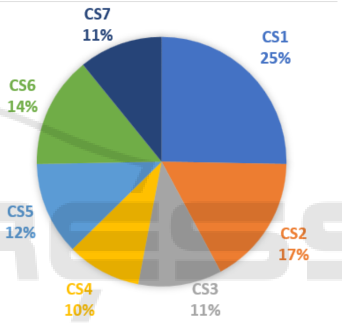


Figure 5: Distribution of concerns per case study.

3.1 Discussion

RQ1: Which Ethical Principles for Developing AI-Based Applications Are Most Often Addressed by Students?
The master's degree study programme of the students whose dissertations we analyse has only one course dedicated to Ethics and Academic Integrity in Data Science. The course teaches general legal and ethical issues of computer science, IP, software licensing (incl. open source), risks and liabilities, privacy, internet and cyberspace, computer security (hacking). There is also some content on ethical AI, such as data access, use, and collection and ethical aspects of re-

Table 1: Fairness and bias principle related concerns.

Concerns	Addr.
CS1: Speed bumps are less common, datasets lack images with them	YES
Object detection uncommon aspect ratio.	YES
Data collection limited to: Black-Yellow Speed Bumps, one small area, one camera.	No
One single annotator bias.	YES
The negative class unclearly defined.	No
”Superior dataset” unbased claim.	No
CS2: Existing annotated dataset split on train/validate/test with no description to ensure relevant evaluation.	No
The classes might be imbalanced or unrepresentative for the context.	No
Model classifies all input as fillings because it is a predominant class.	YES
Dataset not representative for the entire population (gender, age etc.)	No
CS3: Pre-trained tool may contain biases.	No
News content might be unfairly interpreted by sentiment analysis and clustering.	No
Unsupported statement ”impressive efficacy of AI-generated summaries”.	No
CS4: Dataset from MyAnimeList website, with possible user biases (age, region).	No
Classes of anime genres are not balanced.	YES
Some content is never recommended, leading to some artists never being promoted.	No
CS5: Training data is not representative for trading fluctuations (too short).	No
As a Regulated Market for stock trading, this tool should be equally available to all.	No
CS6 : Data distribution on multiple devices can introduce biases, if unbalanced.	No
Dataset has no patient demographics.	No
One class presents fewer samples.	YES
Images generated by other miscroscopic devices might not be used in the model.	No
CS7: Model trained on synthetic data might not be representative.	YES
Access to the app by introducing personal.	No
Marketing user profiles are biased.	No
Popular destinations might become the only recommendations.	No

Table 2: Safety and security principle related concerns.

Concerns	Addr.
CS1: Lack of speed bump standardisation (system won't work in a new region).	YES
Speed bump detection could be used to softnes suspention rather than lower speed.	YES
A false positive can make the car unnecesarily brake, confusing other drivers.	YES
A false negative can lead to dangerous situations if driver is overreliant on the system.	YES
At what distance are speed bumps detected (is there enough time to lower speed)?	No
Most false positives are observed on drainage ditches, followed by crosswalks, but also pole shadows or unpaved roads.	YES
Significant false negatives and low recall suggest speed bump is not detected in time, leading to dangerous situations.	No
CS2: Misdiagnose exacerbates illness.	No
System unable to diagnose less frequent issues, as it's trained only on 4 classes.	No
Diagnose does not replace consultation and specialised treatment.	No
CS3: May increase anxiety, agression or form unbalanced opinions on events.	No
May introduce false elements in summarizing thus spreading misinformation.	No
CS4: Might promote aggressive content.	No
Might promote inappropriate age-content.	No
CS5: Economic decisions are still unpredictable, even with a high model accuracy.	No
User might have financial losses.	No
Artificially create market destabilisation due to uniform predictions and decisions.	
CS6: When can a doctor make safe decisions based on the system input?	No
A misdiagnosys can be dangerous on the life of a pacient.	No
System requires data security for image upload, diagnoses information.	No
CS7: System might recommend places with sudden political conflict (dangerous).	No
System might recommend places with sudden bad weather or natural calamities.	No

search. These aspects of ethical AI from this course seem to be well addressed by the students in their work. In terms of fairness and bias, most students address the issues related to unrepresentative data due to class unbalance, synthetic data, the presence of a predominant class or an under-represented class in the dataset. This is specifically acknowledged in 5 of the 7 case studies and is the most predominant, if not the only aspect related to the principle of fairness and bias that is popular. Data annotation issues are also observed, as they may introduce a significant level of subjective interpretation in the data labelling process if not handled properly. From transparency and explainability, the commonly identified concern is overfitting (in 2 of the 7 papers). In terms of safety and security, only CS1 identifies several cases of possible misuse that can lead to dangerous situations.

As mentioned previously, one of the works (CS1) has identified a higher number of concerns than others. A possible explanation could be the domain of the paper: automotive, a popular subject of interest in industry, that has already attracted many ethical discussions. Given the advanced technology in this area and the high interest from both industry and research, there are many norms and regulations already in place, generated by ethical implications of self-driving car. The effect is that it increases awareness

Table 3: Accountability and liability related concerns.

Concerns	Addr.
CS1: Software system has quality standards to be used in automotive independently, and there are people accountable to fulfil them (maintenance, updates etc.).	No
Responsible party for using the break.	YES
Responsible party for a misclassification is named (software system quality).	No
CS2: Software system quality standards assigned to specific people. The person responsible for a misdiagnosis is specified.	No
CS3: Software system has quality standards assigned to specific people. The person responsible for app malfunction (fake news, discriminating content) is clear.	No
CS4: The user is responsible for the content they to watch.	No
Responsibility for the promotion of biased or aggressive or racist content and its repercussions is assigned to a party.	No
CS5: The user is responsible for their own trading activity and financial losses.	No
Companies may be affected by the software predictions and be sabotaged. The responsibility in this case is clear and software quality ensures no misconduct.	No
CS6: The system is informing the doctor, there are people responsible when the system is at fault (medical device).	No
Software system has quality standards and people accountable to fulfil them, including updates, model maintenance.	No
CS7: User is responsible for choosing a place to travel.	No
The system usage limitations are clear.	No

among the users, developers and researchers alike. This is also perhaps because the safety implications are more direct and more severe than for other topics. Even so, a long-term approach and analysis of ethical implications is necessary in all areas where AI technology is to be frequently used.

In conclusion, most of the students are very familiar with dataset concerns related to dataset containing unbalanced classes. This is probably because they study this aspect specifically in the master's study programme. Also, they are able to identify more ethical concerns in the domains which are highly researched and popular in industry.

RQ2: Which Ethical Principles for Developing AI-Based Applications Are Generally Neglected by Students?

Many ethical concerns are neglected by students in their dissertations. In the area of fairness and bias, there is no concern as to how representative is data to the population (concerning age, ethnicity, gender, etc.), what is the dataset used for the pre-trained models integrated and their biases, or if the data distribution ensures sound evaluation. The resulting societal implications are drastic: some populations will not get a correct dental diagnosis, summarised news content might add more bias and confusion to the user and influence political developments of a region, several young anime artists will never get a chance to get promoted to the public, the stock market can get even more destabilised by speculating AI models, the use of a different microscope for blood cell imaging might get the wrong diagnosis, and several tourist destinations might get overcrowded while others leave people living on tourism without an income. Fairness is hard to achieve even with the best interests at hand, but if not considered at all it may play havoc on peoples lives, their income, safety, or well-being.

Table 4: Transparency and explainability related concerns.

Concerns	Addr.
CS1: Model is black box.	No
Model is overfitting.	No
The user understands the benefits and limitations of the system, how it works, and how to make decisions based on it.	No
Human agency is defined at design level.	No
CS2: Trains a CNN model - black box.	No
Model is overfitting.	YES
Human agency is specified from the design and clear to the user of the app.	No
Results for each class are presented.	No
CS3: The accuracy of tools used is not mentioned (and it often falls under 50%).	No
Uses black box models.	No
It is clear to the user how the system selects balanced and comprehensive set of news.	No
CS4: Black box unexplained.	No
User is clear on what they need to do to get an accurate recommendation.	No
CS5: The model is a combination of network analysis/modelling and decision trees, some decision making explained.	YES
Overfitting is suspected based on results.	No
It is clear how the input features are used in the prediction model.	No
CS6: Black box unexplained.	No
Possible overfitting results.	YES
User understands what images can be used.	No
CS7: Black box unexplained.	No
User is informed on proper app usage.	No

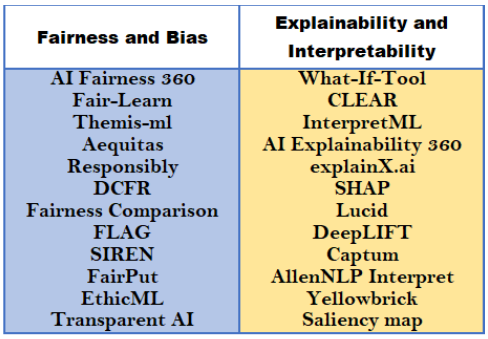
The safety and security risks, as a consequence of system misuse, are also neglected. This is not only related to what an AI model fails to do, but also how that can be exploited by malevolent parties, or simply if the system is used in other ways than intended, and what harm it can do. For example, a false positive diagnosis for a leukemia patient can be equivalent to a sentence to death, while a false diagnose of dental cavity is unlikely to cause a person's death, even if it may damage a tooth. In this context, identifying the severity of the risk of a 0.1% error is very important.

The most neglected ethical principle is accountability and liability, with only one concern addressed of the 13 identified. In this regard, the tendency is to leave the whole responsibility to the user of the AI model. While in many cases the user has an important responsibility of the use of a software system, totally neglecting the developer's responsibility is a huge issue at the moment for the whole software engineering world. If software systems and AI models are supposed to be of help and support to humans, it is crucial that they are reliable tools. Thus, developing systems that we can trust is a responsibility of the developer to ensure the data used, training process, and evaluation are properly performed according to rigorous quality standards. Our AI models can generate very high accuracy, but they might be no good if we can not rely on them, or at least guaranteed a level of quality that makes them a reliable support in certain scenarios. Otherwise, the risk is that our AI models are being used as just toys, and not as a trustworthy decision support technology, always with high caution, and as a result they could become more a hindrance than a human support.

Responsibility in ethical AI development usually derives from transparency and explainability. Although some transparency is addressed in the selected case studies, all students completely neglect the explainability principle in their work. All of the projects involve black box models, meaning we have no knowledge whatsoever on how the model makes its decisions. In consequence, the model might go wrong at any time, and we would not know why. Not knowing the reasoning of an AI models means we can not rely on its 'guesses'. As many explainability tools have been developed lately, we should encourage their use among computer science students of master's degree level, future software developers, and software engineers designing the AI tools of the future.

RQ3: What Solutions for the Implementation of Ethical AI Principles from the Literature Could Be Applied in These Case Studies?
While many of the ethical AI concerns identified deserve a thorough discussion and consideration in a multidisciplinary AI, ethical and law context (Konidena et al., 2024), there are several solutions presented in the literature.

In Figure 6 we present a collection of some representative tools that can be used for ethical AI development. Some are inspired from literature (Cumming et al., 2024), but also online resources such as the Responsible AI Knowledge-base, Git repository (Alexandra, 2023). However, many researchers consider that true accountability come from transparency (Jobin et al., 2019) and that these tools are not sufficient to identify subtle and detailed ethical AI concerns (Bubinger and Dinneen, 2024), and we cannot rely on libraries alone. Ongoing research can help inform developers about which tools to use best for their specific contexts (Jeyakumar et al., 2020), while others militate for developing interpretable AI models rather then trying (rather unsuccessfully) to explain black-box models (Rudin, 2019).



Fairness and Bias	Explainability and Interpretability
AI Fairness 360	What-If-Tool
Fair-Learn	CLEAR
Themis-ml	InterpretML
Aequitas	AI Explainability 360
Responsibly	explainX.ai
DCFR	SHAP
Fairness Comparison	Lucid
FLAG	DeepLIFT
SIREN	Captum
FairPut	AllenNLP Interpret
EthicML	Yellowbrick
Transparent AI	Saliency map

Figure 6: Ethical AI tools.

AI tool development often prioritizes building complex models without considering their industrial fit. Thus, design should start with the intended user and use cases, before implementation, and involve a multi-disciplinary team. Intelligent models should focus on employing interpretable models where possible, specially for high risks decision making. Implementing clear AI Ethics principles in curricula for software quality study is of utmost importance.

3.2 Study Limitations

The number of case studies analysed in this paper is only seven, and can not be sufficiently representative for the general pool of computer science students. Researcher bias could also be present in the interpretations and analysis of the ethical aspects of student dissertations. The study does not aim to perform an exhaustive analysis of ethical AI knowledge, but an assessment used as a starting point for improving the curricula of the master's degree programme. Future work would involve a proper at scale study, perhaps using NLP or LLM - based analysis for automation.

4 CONCLUSIONS

AI applications are crucial for everyday life, but their frequent use raises ethical concerns about data usage and potential misuse. It is crucial for developers to build responsibility and understand the significant impact that the AI technology has on society.

This paper analysed 7 master's degree dissertations in order to assess the level of ethical AI principles addressed by the students, based on four criteria: fairness and bias, safety and security, accountability and liability, and transparency and explainability.

The analysis reveals that the most addressed ethical AI concerns are those related to unbalanced dataset. Explainability is not addressed at all, most works presenting black-box models. The most neglected ethical AI principles are those related to accountability and liability, in which it is expected that the user takes the whole responsibility. Only one of the 7 papers in the study addresses the safety and security concerns of the system developed. Several existing tools for fairness, bias and explainability identified in literature and online resources are recommended both as a support for identifying ethical concerns as well as for mitigation. Other strong recommendations are developing interpretable models and the introduction of ethical AI principles in the curricula of computer science master's degree programme.

ACKNOWLEDGEMENTS

The publication of this article was supported by the 2024 Development Fund of the UBB.

REFERENCES

Alexandra, I. (2023). Responsible ai knowledge-base.

Balasubramaniam, N. (2019). Using ethical guidelines for defining critical quality requirements of ai solutions. Master's thesis, Aalto University.

Bubinger, H. and Dinneen, J. D. (2024). "what could go wrong?": An evaluation of ethical foresight analysis as a tool to identify problems of ai in libraries. *The Journal of Academic Librarianship*, 50(5):102943.

Chiorean, M.-A. (2024). Blood cells classification for acute lymphoblastic leukemia detection using federated learning. Master's thesis, UBB.

Cumming, D., Saurabh, K., Rani, N., and Upadhyay, P. (2024). Towards ai ethics-led sustainability frameworks and toolkits: Review and research agenda. *Journal of Sustainable Finance and Accounting*, 1:100003.

Holgyes, O.-D. (2024). Predicting romanian stock movement trends: a complex network approach combined with machine learning. Master's thesis, UBB.

Horváth, I. (2022). Ai in interpreting: Ethical considerations. *Across Languages and Cultures*, 23(1):1–13.

Huriye, A. Z. (2023). The ethics of artificial intelligence: examining the ethical considerations surrounding the development and use of ai. *American Journal of Technology*, 2(1):37–44.

Jeyakumar, J. V., Noor, J., Cheng, Y.-H., Garcia, L., and Srivastava, M. (2020). How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in neural information processing systems*, 33:4211–4222.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399.

Kenwright, B. (2024). Is it the end of undergraduate dissertations?: Exploring the advantages and challenges of generative ai models in education. In *Generative AI in teaching and learning*, pages 46–65. IGI Global.

Konidena, B. K., Malaiyappan, J. N. A., and Tadimarri, A. (2024). Ethical considerations in the development and deployment of ai systems. *European Journal of Technology*, 8(2):41–53.

Lăzărescu, A.-D. (2024). Anime recommendation system. Master's thesis, UBB.

Miclea, D. (2024). Beyond the headlines: Leveraging ai to streamline news analysis. Master's thesis, UBB.

Moldovan, D.-M. (2024). Advancements in dental care through deep learning. Master's thesis, UBB.

Muntean, A.-O. (2024). Robust speed bump detection based on data collected from gps-enabled dashcams. Master's thesis, UBB.

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., and Nguyen, B.-P. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies*, 28(4):4221–4241.

Remian, D. (2019). Augmenting education: ethical considerations for incorporating artificial intelligence in education. Master's thesis, University of Massachusetts at Boston.

Rosenblatt, M., Dadashkarimi, J., and Scheinost, D. (2023). Gradient-based enhancement attacks in biomedical machine learning. In *Workshop on Clinical Image-Based Procedures*, pages 301–312. Springer.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Tatineni, S. (2019). Ethical considerations in ai and data science: Bias, fairness, and accountability. *International Journal of Information Technology and Management Information Systems (IJITMIS)*, 10(1):11–21.

UNESCO (2024). Ethics of artificial intelligence.

Vakkuri, V. (2022). Implementing ai ethics in software development. *JYU dissertations*.

Wiesenthal, M. (2022). The ethical implications of ai-based mass surveillance tools. Master's thesis, University of Applied Sciences.

Zbârcea, S.-O. (2024). A travel recommandation system based on weather data and traveller profile using machine learning algorithms. Master's thesis, UBB.