AI Systems: Ethical Requirements, Methodological Pitfalls, and Practical Guidelines

TANIA AOUDE, JIHAD DABA
Department of Electrical Engineering,
University of Balamand,
Kalhat, Al-Kurah P.O. box 100 Tripoli,
LEBANON

Abstract: - This article briefly covers the main pitfalls and guidelines for novice researchers in the field of artificial intelligence (AI). The intention is to raise awareness around the existing issues in research to aid in identifying them when conducting a literature survey. While AI offers various benefits, overlooking ethical guidelines and functionality results in major consequences after launching the system. Since publications serve as a reference point for researchers, building on correct techniques allows proper methodology, decreasing the overall chances of unknowingly making major mistakes. The impact of data leakages on the performance of AI systems in the health care domain is further highlighted through a case study of AI for breast cancer detection.

Key-Words: - Artificial intelligence, design ethics, pitfalls, integrity, informative guidelines, benchmarks, trust.

Received: June 16, 2025. Revised: August 29, 2025. Accepted: October 18, 2025. Published: July 7, 2026.

1 Introduction

AI is continuously evolving, surpassing humans in various situations and reinforcing its integration into daily life. This technology is becoming widely used throughout several fields and sectors, impacting people's lives through its advancement. Basic machine learning (ML) systems rarely raised the issue of trust as they simply did what they were told to do based on the coding of their developers. However, their rapid evolution into deeper more complex systems that can operate independently with accuracies that rival that of humans has turned them into a source of fear, [1]. Several concerns arise around the use and development of AI systems with people mentioning privacy, social disadvantages and direct effect on their lives as some of the most worrying. While rules and regulations are set in place for the development of AI products, the focus is mostly on the effect of AI on users rather than the techniques used by developers, [2]. Furthermore, there has been several incidences of AI systems launched for use in the market resulting in terrible consequences such as death, fraud and more, [3], [4]. The latter highlights the fact that market deployed AI technologies are built unsystematically and misleadingly promoted as better than they actually are. Researchers and stakeholders often pay little attention to the actual functionality of their models, focusing mainly on the "ethical" aspect rather than the benefits of the product, [4]. There is a need for awareness around the considerations and guidelines of building and launching AI products, [3]. This article is a beginner oriented guide that focuses on the technical pitfalls, integrity and basis of AI system creation starting with a section on ethical requirements followed by pitfalls and warnings and finally a conclusion.

2 Ethical Requirements

Leading groups have defined sets of procedures that govern AI usage in order to reduce ethical concerns. However, organizations often define their own rules and regulations for ethical development of AI, [5]. A general description of the ethical aspects of AI depends on the quality and practical needs of the stakeholders while adhering to ethical values, [6]. The following sections show the key points involved in the design process.

2.1 Transparency and Explainability

The hidden nature of AI systems makes it difficult to define transparency. However, the main focus is on how the results are obtained. This is why transparency and explainability go hand in hand with the latter reinforcing the clarity of the system by providing insight into the decision making of the model. The general focus is on what and to who to explain, making the requirements bend to the specific task, [5].

2.2 Privacy and Security
Privacy is a basic human right which calls for respect of the secrecy of personal information. This principle is governed by humans through the monitoring of the data acquisition and storage process. Additionally, it ties together with the concept of transparency by requiring a clear reporting of the data handling to safeguard the anonymity of participants, [7].

2.3 Fairness
To classify something as fair means to provide equal opportunities and outcomes for users by creating inclusive products, [7]. This involves making decisions independently of biases throughout the algorithm design process and data selection, [5].

2.4 Accountability
The question of who will take the blame for the resulting outcomes of developed systems raises the issue of accountability. This falls onto the creators and operators of the system to be responsible for the functionality and impact of the product, [7]. Additionally, the division of tasks, such as monitoring and decision making, onto respective parties facilitates the resolution of issues by contacting the specific people concerned, [5].

2.5 Reliability
Reliability is defined differently depending on the sector and organization. For some it is part of the standards regulating safety and quality for risk and purpose assessment. On the other hand, some view it as the root of selecting data that is correct for the intended purpose of the technology being developed, [5].

2.6 Functionality
Very few concentrate on the chance of AI not functioning as originally intended. With the main focus being on the ethical guidelines listed above, a tendency to overlook functionality results in a lack of fulfillment of expectations. This is the result of several failures in the system ranging from engineering faults related to design and implementation all the way to market launching failures that lead to a lack of robustness, [4].

3 Pitfalls and Warnings
Whilst AI developers handle the entire research development phase from beginning to end, their workflow is often under examined. A closer look at the internal workings of these systems reveals crucial mistakes that reduce the reliability of their results. Some leading causes of methodological failures include selecting benchmarks that are not suitable, improper data handling through leakages, misuse of performance metrics and selection biases, [8]. Moreover, the adoption of wrongful technique both in academic frameworks and beyond can have terrible economic and societal impacts. Therefore, it is important to examine whether the developed pipeline is capable of truly predicting new data even if academics rarely involve deployment, [9]. Since it is impossible to mention all types of mistakes in AI systems, the coverage will be focused on the most common issues mentioned at the start of this section.

3.1 Data Leakage
Leakage is referred to as the spread of information into the training set that should be independent of it. Any time a data dependent choice is made in a ML pipeline it must be validated using isolated sets of data. Otherwise, the resulting estimates are overly optimistic and lack generalization capabilities. Some of the ways data can be leaked in a pipeline, especially when using cross validation, are featured below, [9].

3.1.1 Test to Train Leakage
This type of leakage involves seeing some information from the test set in the train set. Special care needs to be taken in cases where the data is dependent on each other to ensure that related samples are not split into separate sets, [9]. Furthermore, certain preprocessing techniques should not be applied to the entire dataset prior to splitting so that the training information is not transformed using statistics from the test set, [10]. A simple experiment was performed to show the effect of this type of leakage on the performance metrics with a pseudocode featured below:

Pseudocode of Trial One and Trial Two

Algorithm 1: Trial one: leakage prevention

```
TrainingData, TestingData = split(Dataset)
AugmentedTrainingData= augment(TrainingData)
model= InitializedModel
model.train(AugmentedTrainingData)
results= model.evaluate(TestingData)
```

Algorithm 2: Trial two: leakage

```
AugmentedData = Augment(Dataset)
TrainingData, TestingData= split(AugmentedData)
model= InitializedModel
model.train(TrainingData)
results= model.evaluate(TestingData)
```

First of all, pretrained VGG16 was selected as the model and an unbalanced breast cancer imaging dataset called MIAS was used as data, [11], [12]. The focus was on the effect of splitting then augmenting versus augmenting then splitting when using the same network and hyperparameters. Trial one split into train, validation and test before augmenting. Trial two augmented and then split into train, validation and test. The results evaluated on the test set are summarized in Table 1 (split then augment) and Table 2 (augment then split).

As shown in the tables, when the dataset was split before augmentation the resulting balanced accuracy was 0.6034. However, when the dataset was augmented and then split, a significantly higher balanced accuracy of 0.8996 was obtained. This shows that leaking data across sets leads to hyper inflated performance and the model wont generalize well to unseen data.

Table 1. Summary of the results of trial one

		Class		
		Benign	Malignant	Normal
Metrics	Precision	0.64	0.43	0.86
	Recall	0.47	0.48	0.94
	F1-score	0.55	0.41	0.89
	Balanced Accuracy	0.6034		

Source: created by the authors

Table 2. Summary of the results of trial two

		Class		
		Benign	Malignant	Normal
Metrics	Precision	0.87	0.89	0.96
	Recall	0.88	0.85	0.97
	F1-score	0.87	0.87	0.96
	Balanced Accuracy	0.8996		

Source: created by the authors

3.1.2 Test to Test Leakage

A form of leakage that occurs when test samples are processed based on information related to other test samples. To give an example, when separate preprocessing workflows are used for the train and test set instead of applying the training data estimations to the test set, [9].

3.1.3 Mitigation Strategies

To reduce the occurrence of these leakages, it is important to describe the steps of the process in great detail. Mentioning the hyperparamters, coding packages, assessment techniques and data handling along with including the code, helps other researchers spot data leakages by attempting to reproduce the work. Additionally, making sure there is proper separation between sets gives a good starting point for avoiding this issue, [9].

Below is a brief list of some precautions to take to ensure leakages are avoided:

1. Selecting the proper splitting criteria for the data depends on the type of data at hand. For example, in the case medical data, it can be split based on: the person which ensures no overlap of the same subject between sets, visit records which is associated with the risk of leakage due to the same subject having repeated visits or after augmentation which has a high possibility of leakage from augmented data based on the same original image ending up in different sets, [13].
2. In the case of time series data in which sample order is meaningful, randomly splitting the data can lead to the model learning from future samples which is why it is recommended to split this type of data chronologically, [14].
3. Applying preprocessing and feature selection based on the training data parameters and then fitting it to the test set, [15].
4. Examining the dataset prior to using it in order to remove duplicate entries, [15].
5. Checking the similarity between the data in the train and test sets to avoid having nearly identical samples between the two, [16].

3.2 Benchmark Selection

The use of benchmarks has turned into a standard for AI models with developers spending large amounts of money on resources in order to get top marks on evaluations. This creates a competitive environment in which creators aim to come out as "winners" even if it requires them to only mention favorable performances while hiding less ideal outcomes, [17]. Several concerns regarding the validity of the performance estimates on benchmarks has resulted in researchers focusing on the misuse and gaps of this evaluation basis, [18]. Certain factors influencing benchmark selection and trust are revealed below.

3.2.1 The Cherry Picking Phenomenon

A top issue in benchmark selection and model assessment is the ability to trick others. This can be easily done given the lack of detail needed to replicate and validate results including absence of code sharing, [18]. Therefore, this hand picking of what to show and what to hide in order to highlight favorable outcomes raises a question on the true validity of these assessments especially when

certain developers have access to more resources making the playing field unfair, [17].

3.2.2 Data Collection

Very few benchmarks specify the source and methodology behind the data collection and documentation process. Additionally, the main research focus is on model development rather than data collection which is a crucial variable for the learning stage. Furthermore, the lack of precision in the annotation of the available data has been found to be a source of skewed results, [18].

3.2.3 Lack of Diversity

The lack of encouragement towards reporting weaknesses is a general issue found in research. However, benchmarking has further amplified this issue as dominant benchmarks being "advertised" by researchers in association with state of the art techniques has caused reuse of select few sets. Furthermore, the data itself lacks representation of minority cases leading to a lack of diversity in the samples themselves, [18].

3.2.4 Suggested Solutions

The political and cultural influence imposed on benchmarks through hype caused by community reporting along with the overall issues that accompany benchmark use, has resulted in calls for unified benchmarks, [17], [18]. Moreover, the absence of proctoring in AI evaluations has caused disparity issues between creators and researchers suggesting a need for regulated control and assessment of their usage, [17].

4 Case Study of AI Based Detection of Breast Cancer

Several studies have examined the effect of data leakages on the performance of AI models for breast cancer detection. Some of which are mentioned below with a focus on performance metric variations.

A study in [19] intentionally leaked features extracted from a pre trained network into their training procedure from the validation set. Subsequently, these extracted features were used to train a separate classifier to distinguish benign and malignant mammography images before reexamining the performance on the validation set. By using the validation set as a reference for selecting the highest performing features the process becomes biased since this information affects the training process. When this final model was tested on a new separate set of data, the overall performance decreased. As the sample sizes of the validation and train sets were decreased this gap became even more apparent. Therefore, when performing feature engineering, it is crucial to ensure that the selected features are in no way dependent on the performance of a set other than the training set to ensure that the training process is not misguided and ungeneralizable to isolated data sets.

Furthermore, the work of [20] conducts a study on several datasets to reveal the over inflation that occurs in model assessment under leakages from preprocessing techniques. Specifically, the breast cancer dataset was used to compare the performance of several well known models between two cases: with leakage and without leakage. While they report that the drop in performance when no leakage is present is only by a slight percentage, the significance of this decrease poses a great impact on the medical domain as it directly affects patients by misclassifying their results. Moreover, some of the models used, such as k nearest neighbors, were found to be capable of resisting data leakages as they maintained high performances with little variation. Given the sensitivity of the results in the medical domain, it is important to have a true estimate of the performance of the models in the real world to reduce over reliance of healthcare professionals on models that appear better than they actually are.

5 Conclusion

In conclusion, digging through the available research and documentation related to AI development has pointed out various gaps and concerns in the field. Both improper methodology and lack of focus towards certain areas has led to the launching and reporting of overly hyped systems. Furthermore, the pool of publications being used by other researchers often fails to transparently detail their pipelines and techniques causing certain issues to go undetected. Since very few start with examining the possible pitfalls associated with ML and simply dive straight into implementation by replicating and enhancing already existing work, they sometimes unknowingly commit mistakes in their methodology. Hence this article serves as a very basic informative guide on some of the most alarming issues in AI development. It is crucial that people interested in working in AI related domains proceed with caution when reading and learning from others. Given the popularity of AI systems in the medical field, the case study of the impact of improper data handling for breast cancer detection

highlights the importance of ensuring proper methodology is used to avoid major consequences.

Since it was not possible to cover everything, it would be of great benefit to extend this work in the future to include more technical explanations and details regarding the misconceptions and fallacies in AI. Additionally, adding more numerical examples to allow the visualization of the impact of improper techniques on model results can strengthen the comprehensibility of this work. Furthermore, providing tabulated summaries of the misconceptions in this field along with their mitigation strategies can be a helpful guide for others provided it is done systematically, is expanded to include the majority of mistakes and includes detailed explanations.

References:

[1] V.Chamola, V.Hassija, A.R.Sulthana, D.Ghosh, D.Dhingra, B.Sikdar, "A Review of Trustworthy and Explainable Artificial Intelligence (XAI)," *Institute of Electrical and Electronics Engineers (IEEE), 2023,* https://doi.org/10.1109/access.2023.3294569.

[2] C. Sanderson et al., "AI ethics principles in practice: Perspectives of designers and developers," *Institute of Electrical and Electronics Engineers (IEEE)*, 2023, https://doi.org/10.1109/tts.2023.3257303.

[3] C. Huang, Z. Zhang, B. Mao, X. Yao, "An overview of artificial intelligence ethics," *Institute of Electrical and Electronics Engineers (IEEE)*, 2023, https://doi.org/10.1109/tai.2022.3194503.

[4] I. D. Raji et al., "The fallacy of AI functionality," *ACM*, 2022, https://doi.org/10.1145/3531146.3533158.

[5] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, S. Kujala, "Transparency and explainability of AI systems: From ethical guidelines to requirements," *Elsevier BV*, 2023, https://doi.org/10.1016/j.infsof.2023.107197.

[6] R. Guizzardi, G. Amaral, G. Guizzardi, J. Mylopoulos, John," Ethical requirements for AI systems", *33rd Canadian Conference on Artificial Intelligence*, 2020, https://doi.org/10.1007/978-3-030-47358-7_24.

[7] P. Brey and B. Dainow, "Ethics by design for artificial intelligence," *Springer Science and Business Media LLC*, 2023, https://doi.org/10.1007/s43681-023-00330-4.

[8] Z. Luo, A. Kasirzadeh, and N. B. Shah, "The more you automate, the less you see: Hidden pitfalls of AI scientist systems," 2025, https://doi.org/10.48550/arXiv.2509.08713.

[9] L. Sasse et al., "On leakage in machine learning pipelines," 2024, https://doi.org/10.48550/arXiv.2311.04179.

[10] A. Apicella, F. Isgrò, and R. Prevete, "Don't push the button! exploring data leakage risks in machine learning and transfer learning," *Springer Science and Business Media LLC*, 2025 https://doi.org/10.1007/s10462-025-11326-3.

[11] K. Simonyan, A. Zisserman," Very deep convolutional networks for largescale image recognition" *ArXiv preprint arXiv,* April 2015, https://doi.org/10.48550/arXiv.1409.1556.

[12] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, P. Taylor, D. Betal, J. Savage, "Mammographic Image Analysis Society (MIAS) database v1.21", *Apollo - University of Cambridge Repository*, August 2015, https://doi.org/10.17863/CAM.105113.

[13] D. J. Rumala, "How you split matters: Data leakage and subject characteristics studies in longitudinal brain MRI analysis," 2023, https://doi.org/10.48550/arXiv.2309.00350.

[14] M. A. Lones, "Avoiding common machine learning pitfalls," 2024, https://doi.org/10.1016/j.patter.2024.101046.

[15] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in ML-based science," 2022, https://doi.org/10.48550/arXiv.2207.07048.

[16] M. Gupta, J. Christopher, and K. Ramisetty, "Mitigation strategies for data leakage in machine learning system design," *Institute of Electrical and Electronics Engineers (IEEE),* 2025, https://doi.org/10.1109/TENSYMP63728.2025.11144995.

[17] Z. Cheng, S. Wohnig, R. Gupta, S. Alam, T. Abdullahi, J. Alves, C. Nielsen-Garcia, S. Mir, S. Li, J. Orender, S. A. Bahrainian, D. Kirste, A. Gokaslan, M. Glinka, C. Eickhoff, and R. Wolff, "Benchmarking is broken - don't let AI be its own judge," *Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing System Processing* (*NeurIPS 2025*), Ssan Diego, CA, 2025, https://doi.org/10.48550/arXiv.2510.07575.

[18] M. Eriksson, E. Purificato, A. Noroozian, J. Vinagre, G. Chaslot, E. Gómez, and D. Fernandez-Llorca, "Can we trust AI benchmarks? an interdisciplinary review of current issues in AI evaluation," *Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society* (*AIES-25*), Madrid, Spain, 2025, https://doi.org/10.1609/aies.v8i1.36595.

[19] R. K. Samala, H. Chan, L. Hadjiiski,S. Koneru, "Hazards of data leakage in machine learning: A study on classification of breast cancer using deep neural networks," *SPIE* , 2020, https://doi.org/10.1117/12.2549313.

[20] M. A. Bouke, S. A. Zaid, A. Abdullah, "Implications of data leakage in machine learning preprocessing: A multi-domain investigation," *Research Square*, 2024, https://doi.org/10.21203/rs.3.rs-4579465/v1.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.