Pollution with Purpose: The Role of Data Quality in Trustworthy AI

Leonie Louisa Etzold[1,*], Tim Robin Kosack[1], Oscar Hernán Ramírez-Agudelo[1], Clemens Danda[1] and Michael Karl[1]

[1]German Aerospace Center, Institute for AI Safety and Security, Rathausallee 12, 53757 Sankt Augustin, Germany

Abstract

This paper contributes to the practical implementation of the third pillar of trustworthy AI: technical and social robustness. We enhance the robustness and reproducibility of an AI model through the integration of polluted data into the AI training process. To do so, the YOLO11n model backbone is fine-tuned with a subset of the PASCAL VOC12 benchmark data using different shares and intensities of a horizontal blur polluter in the training data. Through this approach we are able to reach a significant increase in robustness on similarly polluted test data. Hereby, the training of AI systems becomes better aligned with context- and environment-specific conditions. This approach does not only contribute to the technical robustness of AI systems but poses the opportunity to also boost their social robustness, by increasing their adaptability to diverse and dynamic real-world settings.

Keywords

robust AI, data quality, trustworthy AI, data corruption

1. Introduction

Applying and assessing trustworthy AI requires (1) an AI-focused definition or concept of trustworthiness, and (2) an approach for the operationalization of 'trustworthiness'. Only if both aspects are combined, statements about the trustworthiness of an AI system can be made.

To this end, the European Commission's High-Level Expert Group (HLEG) developed the *Ethics Guidelines for Trustworthy AI* (2019), as well as their approach for the proper operationalization, the *Assessment List for Trustworthy AI* (ALTAI). According to the HLEG, trustworthy AI has three central components: It should be lawful, ethical, and robust [1].

In this paper, we discuss a promising approach to contribute to the third pillar of trustworthy AI robustness, which must be considered from both a technical and social perspective [1]. We enhance the robustness and reproducibility of an AI model through the integration of polluted data into the AI training process. Inappropriate training data during the development may result in invalid outcome once the model is exposed to real-world inference data. This would decrease the technical robustness of the AI system and, accordingly, reduce its overall trustworthiness. Thus, we underline the importance of Data Quality (DQ) regarding the robustness of an AI model. Bringing the AI model's training data closer to the data to which it is exposed in the real world may reduce the its susceptibility to polluted data in certain situations.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.

leonie.etzold@dlr.de (L. L. Etzold); tim.kosack@dlr.de (T. R. Kosack); oscar.ramirezagudelo@dlr.de (O. H. Ramírez-Agudelo); clemens.danda@dlr.de (C. Danda); michael.karl@dlr.de (M. Karl)

0009-0001-3442-3242 (L. L. Etzold); 0009-0006-5101-7541 (T. R. Kosack); 0000-0002-9379-5409 (O. H. Ramírez-Agudelo); 0000-0002-7785-7673 (C. Danda); 0009-0001-0535-0515 (M. Karl)

2. Background and Related Work

2.1. Data Quality

While both the *Ethics Guidelines for Trustworthy AI* as well as the ALTAI are more focused on ethics than engineering, some requirements allow the derivation of technical approaches for ethical problems [2]. Furthermore, most ethical requirements of the ALTAI are in some way connected to the technical aspect of the quality of data being used during the AI life cycle: DQ is not only connected to robustness and reproducibility, but also to key-issues of discrimination [3], privacy [4], and sustainability [5]. The EU Artificial Intelligence Act ('AIA') mandates that data quality and robustness be ensured in high-risk AI systems (Art. 10 and 15 AIA). This specifically concerns the application of AI systems in contexts such as social categorization or critical infrastructure – domains identified as carrying significant risks of human rights infringement (Annex III). Specifically, the AI Act mandates that training, validation, and testing datasets used in high-risk AI systems be relevant, representative, and to the best extent possible free of errors. During the legislative process, the wording was somewhat relaxed, limiting responsibility for residual risks that cannot reasonably be controlled ('best extent possible'). Regarding robustness, Art. 15 AIA requires that high-risk AI systems are 'as resilient as possible' achieved through technical means such as redundancy solutions. This paper highlights specific technical aspects to strengthen robustness by developing 'appropriate technical solutions to prevent or minimise harmful or otherwise undesirable behaviour' (Rec. 75 AIA).

This paper is concerned with the impact of deteriorating DQ on the robustness of AI perception models. ISO 5259:2024 [6] defines DQ as a characteristic which describes whether the data are in compliance with the requirements set for a specified context. In the context of AI development, the quality requirement can be defined as a prerequisite which needs to be met for the AI model to achieve pre-defined quality properties (such as levels of performance). Inherent to these definitions is the assumption that the quality of data affects the performance of AI models, which has been investigated and affirmed thoroughly over the previous years [7, 8, 9]. As a result, the paradigm of data-centric AI emerged postulating an intensified focus on the quality of data used for machine learning purposes [10, 11].

It is to be differentiated between data used during AI development and the new and unlabelled inference data , the model is met with during usage. The larger the gap between these types of data is, the more the performance of AI models suffers. However, for the development data to be completely representative of every possible manifestation of inference data is hardly feasible due to the complexity of real-world events. Besides differences in distribution between data sets or temporal evolution due to the non-stationarity of data [12, 13], differences between the quality of development and inference data challenge the performance of AI models in application.

As a method to mitigate this, corruption-based data augmentation is increasingly used as a standard for safety-critical applications by training models with (partially) polluted data. Hereby, the model can achieve increased robustness towards the type(s) of pollution it was trained with.

2.2. Robustness

Social robustness of AI systems can be understood as their robustness to (unpredictable) social changes. This may include erroneous or unseen data (e.g., incorrect data collection in marginalized groups) [14] and shifted perspectives or goals (e.g., political goals or social perception). In this regard, acceptance of AI systems and social robustness are interdependent: AI systems which are unable to adapt to new (social) circumstances will most likely not be accepted. While this broad definition is highly dependent on the respective context, the technical notion of robustness provides measures which reflect this adaptation.

Technical robustness provides a measure on the resilience of AI models towards sub-optimal application environments. In this regard, AI robustness "refers to the ability of a system to maintain consistent performance when exposed to diverse, unexpected or even adversarial inputs" [15]. Thus, robust AI

systems are resilient to noise, can perform with a variety of datasets, resist to adversarial attacks and are both interpretable and explainable [15]. In this paper, we focus on the resilience to noise.

Formally, Zhang et al. [16] define the technical robustness of an AI model as a measurement of the difference between the correctness of the output of the model under optimal versus perturbed conditions. When AI is applied in use cases which are time-critical, such as automated driving or disaster management, the inference data is more prone to issues in data quality since the time available for data collection and processing is limited during the usage phase. AI models which were developed under optimal conditions with regard to data quality will lack robustness towards deteriorating inference data quality [7]. To investigate the robustness of AI models towards deteriorating data quality in inference data, test data is polluted to imitate potentially occurring issues in inference data.

We adapt Zhang et al.'s quantification of robustness [16] for our experiments as

$$r = E(M) - E(\delta(M)) \tag{1}$$

with $E(M)$ as the level of correctness of AI model M and $\delta(M)$ signifying the AI model with perturbations on inference data. To allow for a standardization of the interpretation of r, we redesign the formula as follows

$$r = 1 - (E(M) - E(\delta(M)). \tag{2}$$

As a measure of the correctness E of the AI model, we choose the mean Average Precision (mAP), i.e. the mean value of average precision over all classes. More specifically, the mAP averaged over intersection over union (IoU) thresholds from 0.50 to 0.95 in steps of 0.05 (*mAP50:95*) is chosen which was first introduced within the COCO documentation. It allows for a broad assessment of the model's performance balancing recall and precision. [17, 18]

Therefore, we define robustness as

$$r = 1 - (mAP@0.50:0.95^{clean} - mAP@0.50:0.95^{polluted}) \tag{3}$$

with $mAP@0.50:0.95^{clean}$ displaying the mean average precision on clean inference data and $mAP@0.50:0.95^{polluted}$ on polluted inference data. Since the *mAP@0.50:0.95* is limited to values between 0 and 1, 1 signifying the highest possible precision, our measure of robustness r signals low robustness to increased DQ issues in inference data when approaching 0 and high robustness if it approaches 1. Once r exceeds 1, this signifies a better performance on polluted data compared to optimal conditions with the gap in performance increasing with r approaching 2. Depending on the model type and context of application, this equation works with any other performance indicator that can be normalized to values between 0 and 1 with convergence to 1 indicating high performance. We developed this notion of robustness to specifically consider machine learning models and it can be adapted to other metrics of performance which can be normalized to values between 0 and 1, increasing with performance. An in-depth comparison to other notions of robustness will be investigated in future work.

Since we are assessing the performance of each model on a multitude of pollution intensities, robustness towards multiple intensities of pollution will be estimated by the application of an average over the performance of the models on the different test intensities. In this case, the correctness of the model with decreased quality of different intensities is defined as

$$\mu mAP@0.50:0.95^{polluted} = \frac{\sum_{p=1}^{P} mAP@0.50:0.95_{p}^{polluted}}{P} \tag{4}$$

with P signifying the number of different pollution intensities. For completion, the final formalization for average robustness is

$$\mu r = 1 - (mAP@0.50:0.95^{clean} - \frac{\sum_{n=1}^{P} mAP@0.50:0.95_{p}^{polluted}}{P}) \tag{5}$$

or

$$\mu r = \frac{1 - (mAP@0.50:0.95^{clean} - \sum_{p=1}^{P} mAP@0.50:0.95_p^{polluted})}{P}. \quad (6)$$

However, it is not possible to draw conclusions from neither r nor μr on the level of performance of the investigated models. Therefore, for a complete assessment, both $mAP@0.50:0.95^{clean}$ and $mAP@0.50:0.95^{polluted}$ or $\mu mAP@0.50:0.95^{polluted}$ should be taken into account, respectively.

The empirical work conducted in this paper adds to the previously published literature [19, 20, 21] as it provides evidence that the robustness of an AI system can be enhanced by integrating deliberately perturbed (or "polluted") data into the training process.

3. Methodology

3.1. Data

The benchmark data set chosen for analysis is the PASCAL VOC12 data set [22]. It consists of 11,530 images containing 27,450 annotated objects from 20 classes from the categories *Person*, *Animal*, *Vehicle*, and *Indoor*. PASCAL VOC12 was chosen due to its similarity to COCO which is the foundation for the YOLO object detection models as described in Section 3.2. It is widely used for AI assessment within the literature [23, 24]. The objects within PASCAL VOC12 are marked as *difficult*, *truncated*, or *occluded*, if applicable. For the means of this analysis, the data set was subset to only include images that display ground-based vehicles, i.e. bicycles, busses, cars, motorbikes, and trains. While the data set was subset by the previously stated classes, the class *Person* was included additionally for training and inference. Furthermore, objects that were marked as difficult were excluded, while objects marked as truncated and occluded remain in the data set.

Table 1
Comparison of PASCAL VOC12 original data set to custom data set used in the following work.

	Original	Subset
Number of Classes	20	6
Number of Images	11,530	3,055
Annotated Objects	27,450	7,504
Object Groups	Person, Animal, Vehicle, Indoor	Person, Vehicle
Included difficulties	Difficult, truncated, occluded	Truncated, occluded

The final data set consists of 2,055 images, including 7,504 objects from six classes.[1] A direct comparison between the original and subset data set is displayed in Table 1 and class distribution is displayed in Table 2. The images in the data set were resized such that the longest image side measures 416 pixels to obtain comparable pollution intensities across all images.

Table 2
Class distribution in custom data set.

Bicycle	Bus	Car	Motorbike	Person	Train
753	638	2,105	763	2,573	672

[1]Our code including the creation of the data set and all experiments will be made available online pending institute permission. To get access to the repository, please contact the authors.

3.2. Model

We chose Ultralytics YOLO11 [25] as a state-of-the-art framework which is highly utilized for real-time object detection across various domains [26, 27]. YOLO is widely used in the literature in research on object detection and transfer learning [23, 28, 29]. Specifically, we are using YOLO11n as a pre-trained model to fine-tune with our data set. YOLO11n is trained with the COCO data [30] set consisting of 80 classes, including ground-based vehicles and persons, rendering it fitting for the observed fictitious use-case.

3.3. Data Pollution

As a polluter to test our hypotheses with, we choose a horizontal blur to mimic motion blur pollution, which may occur when data collection includes either moving cameras or moving targets. We apply a square convolutional kernel which averages over the horizontally neighbouring pixels. Following [31], it can be formalized as

$$box_{N,N}[n,m] = \begin{cases} \frac{1}{k} & \text{if } -N \leq n \leq N \text{ and } m = 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

with k as the kernel width $2N+1$ and for N in the range of 1 to 22 in steps of 3, resulting in the kernel width k ranging from 3 to 45 pixels in steps of 6. For kernel width k, a pixel of input image I at position x, y is convoluted to output pixel $I'(x,y)$ as follows

$$I'(x,y) = \frac{1}{k} \sum_{n=-N}^{N} I(x+n, y). \tag{8}$$

Following, the blur intensities will be expressed as the percentage of kernel width to maximum image side length of 416 pixels, i.e. as $k/416\%$. This leads to a range of intensities of 0.72 to 10.82% in steps of 1.5%. Figure 1 displays an example image with no blur as well as intensities of 0.72, 6.49, and 10.82%. While the image with low the lowest blur intensity is only marginally perturbed, the higher blur intensities display a high level of pollution.

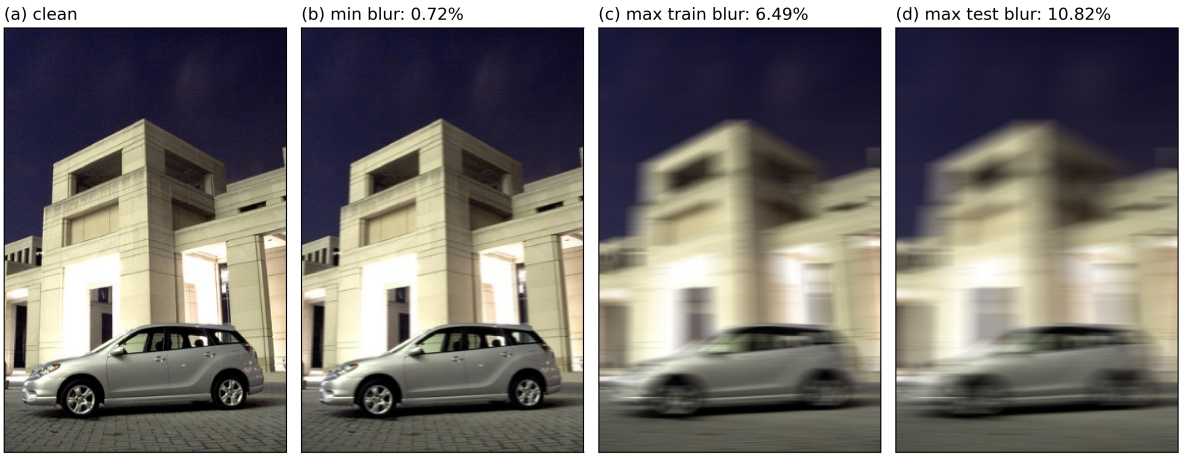

Figure 1: Example image with (a) no blur pollution, (b) minimal blur pollution, (c) maximum blur pollution used for training, (d) maximum blur pollution used for testing.

3.4. Model Training and Testing

Transfer learning is used to train the model, i.e. the fine-tuning process with the data set described in Section 3.1 is conducted using the back-bone of YOLO11n. For each experiment, the model set up

is held constant using the parameters in Table 3. For testing, the best model is chosen for each of the experiments.

Table 3
Model training parameters.

Parameter	Value
Epochs/Patience	75/25
Image Size	416
Pre-Trained	Yes
Optimizer	SGD
Starting learning rate	0.005

A split ratio of 70% (1,439 images) for training and both 15% (308) for validation and test data was applied and implemented using multi-label stratified split to ensure similar distribution across all data sets. Each experiment including the splitting process was repeated over 10 seeds to mitigate the possibility of randomly occurring bias in the distribution.

Table 4
Experimental parameters.

Experimental Parameter	Values
Training & validation pollution shares	0%, 10% 20%, 40%, 60%, 80%, 100%
Test pollution shares	0%, 100%
Training & validation pollution intensities	0.72%, 2.16%, 3.61%, 5.05%, 6.49%
Test pollution intensities	0.72%, 2.16%, 3.61%, 5.05%, 6.49%, 7.93%, 9.36%, 10.82%
Seeds	1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Within one seed, the model was trained using the same training and validation splits which were varied in (1) the share of polluted data, and (2) the intensity of pollution. For each split, a baseline model was trained on clean training and validation data. Table 4 displays the full list of experimental parameters and values. The additional models were trained using shares between 10 and 100% of polluted training and validation data and blur intensities between 0.72 and 6.49%. Each model was then tested on a clean test set as well as test sets fully polluted with blur intensities of 7.93, 9.36, and 10.82% additionally to those used for training.

3.5. Ongoing Experiments on Additional Pollution Types

To validate the experiments conducted in this paper, we are currently working on additional experiments investigating further pollution types. While first results on experiments on Gaussian noise appear to corroborate the results reported in the following section, the current status of the experiments is not advanced enough to report on them.

4. Results

Figure 2 displays the performance of the different models on increasingly polluted test data for each share of training pollution. The grey line corresponding to a training blur of 0% is trained on clean data. It therefore acts as the baseline model and does not vary across the different shares (a) to (f). For clean test data, the baseline model reaches a mAP of 0.92, indicating a good model performance under unperturbed conditions. However, its decreases rapidly with increasing test pollution. While its mAP comes within 0.5 for the first time at a blur intensity of 3.6%, it falls to 0.042 at the maximum pollution of 10.8%. The models with the smallest intensity of blur in the data mirrors the decline of the clean model, however the performance is slightly better for polluted test blur, with the performance on polluted

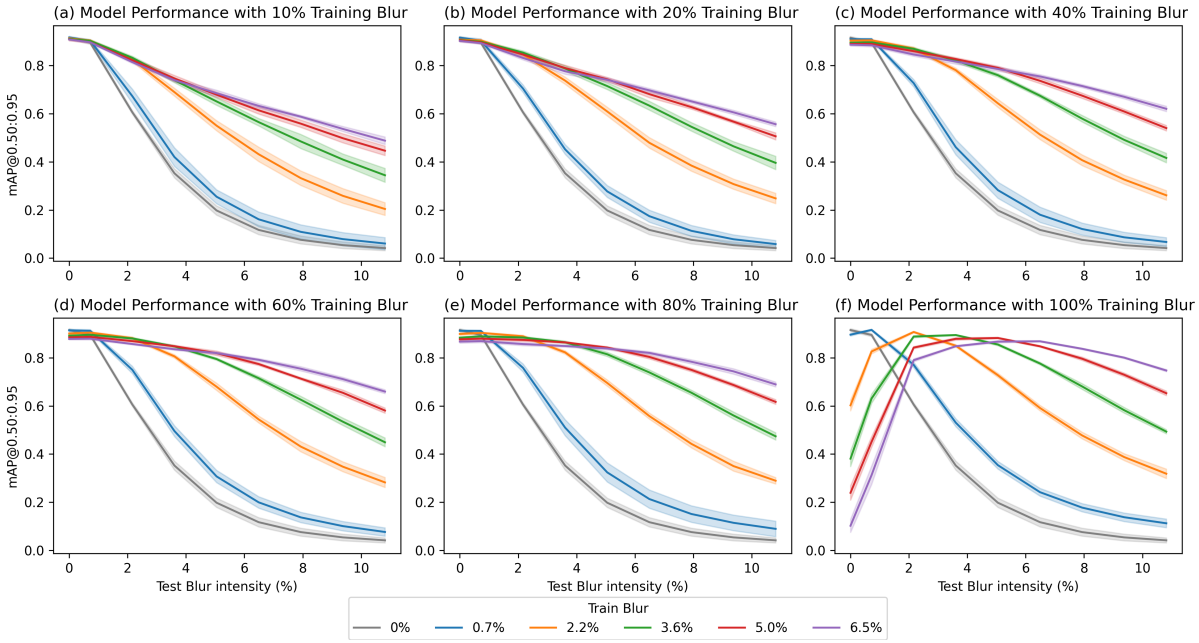

Figure 2: The evolution of mean average precision mAP@0.50:0.95 of YOLO object detection models on test data with different intensities (0.7, 2.2, 3.6, 5.0, 6.5, 7.9, 9.4, 10.8%), with training and validation data varying in intensity of pollution (0.7, 2.2, 3.6, 5.0, 6.5%) and shares of polluted training and validation data of 10, 20, 40, 60, 80, and 100% for graphs (a) to (f) respectively.

test blur increasing with the share of polluted training data. Models with training intensities of 2.2 to 6.5% show more pronounced differences compared to baseline, even at a low share of polluted training data of 10%. In Figures 2 (a) to (e), the decrease in performance is much lower with increasing test blur intensities than it is for baseline, with both higher shares and intensities of polluted training data reducing the overall decrease in performance. While higher shares and intensities of polluted training data decrease the performance on clean data slightly, the advantage polluted models have on polluted test data is distinctly higher. For the model with 80% of training data polluted at the highest level of 6.5%, the decrease in mAP amounts to 0.048 compared to baseline on clean data, while it is by 0.648 higher than baseline at a test blur intensity of 10.8%. An interesting observation to be made is that even though each model displayed in Figures 2 (a) to (e) is only familiar with clean data and data polluted with the respective training intensity, the performance of the model also improves significantly on test data of pollution intensities which are unfamiliar to the model. All models perform as well or better on test data which is polluted in a lesser intensity than the model is trained with. This changes for Figure 2 (f) which displays the performance of models trained on fully polluted data. These models were not made familiar with clean data during the fine-tuning process and generalize poorly to clean test data while they also do not perform as well towards data with a lesser pollution intensity, resulting in them peaking in performance on the pollution level their training data was polluted with. This underscores an observation which holds true for all graphs in this figure: For each test intensity, the model trained with the same pollution intensity reaches the highest performance within one training share. This effect increases in pronunciation with increasing training share.

To allow for the integration of these levels of performance into the concept of robustness as previously defined in Section 2.2, Figure 3 displays the average robustness and their components. Figure 3 (a) shows the components of Equation (3) for each of the models. At a share of 0%, the models were trained on clean data showing the performance of the baseline model. The mAP on clean data ($mAP@0.50:0.95^{clean}$) indicates the performance of the trained models on clean test data. The models trained with the lowest blur intensity of 0.7% remain almost constant in performance over the different shares of polluted training data. It even increases performance slightly when introduced at lower levels. For higher train blur intensities, increasing shares in polluted training data lead to a monotonous decline in performance

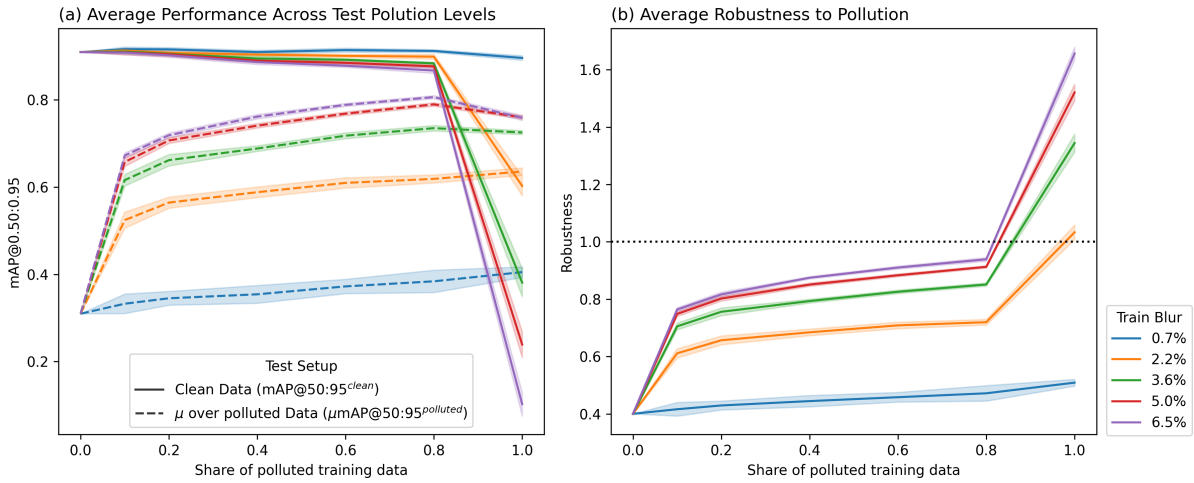

Figure 3: (a) The mean average precision mAP@0.50:0.95 on clean (solid) and the average over all polluted (dashed) test data sets for models trained on different shares and intensities of polluted training and validation data. (b) The corresponding average robustness of these models as defined in equations 5 and 6, with the dashed line corresponding to the tipping point to reversed robustness, $r = 1$.

up to a share of 80%. The decline increases with the severity of pollution intensity and ranges from a minimal decline of 0.003 for a training blur pollution of 0.7% to 0.048 for 6.5% compared to the baseline performance of 0.92. The decrease is rather small, suggesting that generalization to clean data can still be obtained to a large degree. However, at 100% polluted training data, average precision decreases significantly for pollution exceeding the intensity of 0.7%. While the lightest pollution intensity experiences only a small decline of 0.019, the higher training intensities increasingly plummet with the 6.5% intensity dropping the maximum amount of 0.813. The average mAP over polluted data ($\mu mAP@0.50:0.95^{polluted}$) show an average of the model performance on the different test pollution intensities as expressed in Equation (4). For intensities above 0.7%, the highest increase in average performance can be observed between clean training data and the lowest share of polluted training data of 10%. From there on, average performance keeps on increasing moderately up to a training share of 80%. For shares up to 80% of polluted training data, all models perform better on clean data compared to the models' average performance on polluted data. However, at 100% polluted test data, performance on clean test data drops below that on polluted data for all intensities but 0.7%.

Figure 3 (b) displays the full measures of robustness as defined in Equations (5) and 6. It is apparent that the average robustness increases both with intensity of training blur and share of polluted training data, while the measures for average robustness converge with higher intensities. This increase is, again, least pronounced for the lowest training blur intensity. At fully polluted training data, all models from training intensity 2.2% and up exceed the mark of 1, indicating that the performance of these models are worse on clean than on polluted data. We can observe the least robust model to be the baseline model with an average robustness of 0.405. This indicates a gap between performance on clean data and average performance on polluted data of 0.595. The highest level of average robustness without reversing can be observed at training with a share of 80% of polluted training data at the highest intensity of 6.5%. Here, the average robustness amounts to 0.939, indicating a gap between performance on clean and polluted data of 0.061. However, this does not only come from the increase in performance on polluted data, but also on a decrease on clean data of 0.042 compared to baseline.

5. Discussion

This paper's results indicate a significant increase in robustness towards DQ deterioration in inference data when models are trained with data that was perturbed with the corresponding pollution, in this case horizontal blur. This adds to previous findings indicating that data which the model is made

familiar with during training is more easily identified during testing, even if the data undergoes a loss of information. Our findings indicate that robustness increases with the share and intensity of polluted training data. However, the trade-off between decrease of performance on clean data and increase on polluted data increases with the same parameters. Therefore, it appears to be advised to balance both the share and intensity of polluted training data based on observed or expected pollution in real-world inference data to prevent a potentially unnecessary decrease in model performance on clean data. Thus, the added value is highly dependent on the specific use case.

An in-depth evaluation of both development and inference data is further advised to truly increase trustworthiness through robustness from both a technical and social perspective. To do so, potential quality issues within data must be identified dependent on the use case to subsequently introduce relevant data corruptions in the training data.

6. Conclusion and Future Work

This paper examined the interrelation between data quality, robustness, and the trustworthiness of AI systems. We demonstrated that the robustness of an AI system can be enhanced by integrating deliberately perturbed data into the training process. Hereby, the training of AI systems becomes better aligned with context- and environment-specific conditions. This approach does not only contribute to the technical robustness of AI systems but poses the opportunity to also boost their social robustness, by increasing their adaptability to diverse and dynamic real-world settings. Incorporating new or previously underrepresented features into data sets offers the opportunity to better capture complex social nuances, contributing to more context-aware and socially responsive AI systems. Enhancing both technical and social robustness is closely linked to the third pillar of Trustworthy AI [1]. While improving the trustworthiness of AI systems is generally advantageous, it is particularly critical in unpredictable high-risk scenarios such as floods, pandemics, or wildfires.

There are some limitations to this work at this point in time which we want to address. First of all, the results discussed correspond to the YOLO11n object detection model and need to be contextualised with further research on additional model specifications. Furthermore, the perturbation type is limited both in scope and transferability to real world blur. Future work may therefore focus on the identification and as close as possible imitation of perturbations observed in real-world data and be tested on data that displays real-world blur for validation purposes. Investigating whether this approach also facilitates increased robustness to different types of pollution within one model is promising.

Also, the applicability of this approach highly depends on the context within which the AI model is used. Choosing suitable pollution types as well as intensities requires use-case specific investigation of the sensors, expected environments and comparable data. Future research should interest itself with the formalization of this process, whereas we expect a combination of both ex-ante and iterative approaches to yield most promising results. Therefore, an in-depth focus on the use-case specific application is essential.

Finally, while we see this paper as a contribution to the third pillar of trustworthy AI, DQ is always connected to ethical (e.g., data acquisition, completeness) and legal (e.g., data protection, AI Act) aspects. Especially, if we want to have technical and social robustness as trust-enablers, it might be necessary to add additional safeguards like (trained) human operators. Further exploration of this relation is essential to deepen our understanding and move closer to realizing the vision of trustworthy AI.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o in order to: Generate Literature Review and Aid Programming. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] European Commission and Directorate-General for Communications Networks, Content and Technology, Ethics guidelines for trustworthy AI, Publications Office, 2019. doi:doi/10.2759/346720.

[2] T. Stefani, F. Deligiannaki, C. Berro, M. Jameel, R. Hunger, C. Bruder, T. Krüger, Applying the assessment list for trustworthy artificial intelligence on the development of ai supported air traffic controller operations, in: 2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC), 2023, pp. 1–9. doi:10.1109/DASC58513.2023.10311323.

[3] S. Moussawi, X. N. Deng, K. D. Joshi, AI and Discrimination: Sources of Algorithmic Biases, SIGMIS Database 55 (2024) 6–11. URL: https://doi.org/10.1145/3701613.3701615. doi:10.1145/3701613.3701615.

[4] M. Giuffrè, D. L. Shung, Harnessing the power of synthetic data in healthcare: innovation, application, and privacy, NPJ digital medicine 6 (2023) 186.

[5] Y. Li, X. Chao, Toward sustainability: Trade-off between data quality and quantity in crop pest recognition, Frontiers in Plant Science Volume 12 - 2021 (2021). URL: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2021.811241. doi:10.3389/fpls.2021.811241.

[6] International Organization for Standardization (ISO), ISO/IEC 5259-1:2024: Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples , 2024. URL: https://www.iso.org/standard/81088.html.

[7] T. He, S. Yu, Z. Wang, J. Li, Z. Chen, From data quality to model quality: an exploratory study on deep learning, 2019. doi:10.48550/arXiv.1906.11882.

[8] Q. Liu, W. Ma, Navigating data corruption in machine learning: Balancing quality, quantity, and imputation strategies, 2024. doi:10.48550/arXiv.2412.18296.

[9] G. Fenza, M. Gallo, V. Loia, F. Orciuoli, E. Herrera-Viedma, Data set quality in machine learning: Consistency measure based on group decision making, Applied Soft Computing 106 (2021) 107366. doi:10.1016/j.asoc.2021.107366.

[10] J. Jakubik, M. Vössing, N. Kühl, J. Walk, G. Satzger, Data-centric artificial intelligence, Business & Information Systems Engineering (2024). doi:10.1007/s12599-024-00857-8.

[11] M. H. Jarrahi, A. Memariani, S. Guha, The principles of data-centric ai, Communications of the ACM 66 (2023) 84–92. doi:10.1145/3571724.

[12] E. Han, C. Huang, K. Wang, Model assessment and selection under temporal distribution shift, 2024. doi:10.48550/arXiv.2402.08672.

[13] W. Fan, P. Wang, D. Wang, D. Wang, Y. Zhou, Y. Fu, Dish-ts: A general paradigm for alleviating distribution shift in time series forecasting, 2023. doi:10.48550/arXiv.2302.14829.

[14] B. chander, C. John, L. Warrier, K. Gopalakrishnan, Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness, ACM Comput. Surv. 57 (2025). URL: https://doi.org/10.1145/3675392. doi:10.1145/3675392.

[15] P. Roy, Enhancing real-world robustness in ai: Challenges and solutions, JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING 12 (2024) 34–49. doi:10.70589/JRTCSE.2024.1.6.

[16] J. M. Zhang, M. Harman, L. Ma, Y. Liu, Machine learning testing: Survey, landscapes and horizons, 2019. doi:10.48550/arXiv.1906.10742.

[17] COCO Consortium, Coco api, 2015. URL: https://github.com/cocodataset/cocoapi.

[18] Ultralytics, Ultralytics yolo documentation: Performance metrics deep dive, 2023. URL: https://docs.ultralytics.com/guides/yolo-performance-metrics/.

[19] V. Karpukhin, O. Levy, J. Eisenstein, M. Ghazvininejad, Training on synthetic noise improves robustness to natural noise in machine translation, 2019. doi:10.48550/arXiv.1902.01509.

[20] I. Vasiljevic, A. Chakrabarti, G. Shakhnarovich, Examining the impact of blur on recognition by convolutional networks, 2016. doi:10.48550/arXiv.1611.05760.

[21] D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, 2019. doi:10.48550/arXiv.1903.12261.

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html, 2012.

[23] N. Zhao, Enhancing object detection with yolov8 transfer learning: A voc2012 dataset study, in: Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence, SCITEPRESS - Science and Technology Publications, 2024, pp. 429–434. doi:10.5220/0012939600004508.

[24] Q. Zhou, C. Yu, Z. Wang, Q. Qian, H. Li, Instant-teaching: An end-to-end semi-supervised object detection framework, 2021. doi:10.48550/arXiv.2103.11402.

[25] G. Jocher, J. Qiu, Ultralytics YOLO11 (Version 11.0.0, License AGPL-3.0) [software], 2024. URL: https://github.com/ultralytics/ultralytics.

[26] L. HRIC, J. A. BACIK, D. PERDUKOVA, Real-time object detection of simple drawings using yolo11 on constrained datasets, MM Science Journal 2025 (2025). doi:10.17973/mmsj.2025_03_2024118.

[27] M. Mao, M. Hong, Yolo object detection for real-time fabric defect inspection in the textile industry: A review of yolov1 to yolov11, Sensors (Basel, Switzerland) 25 (2025). doi:10.3390/s25072270.

[28] L.-H. He, Y.-Z. Zhou, L. Liu, W. Cao, J.-H. Ma, Research on object detection and recognition in remote sensing images based on yolov11, Scientific reports 15 (2025) 14032. doi:10.1038/s41598-025-96314-x.

[29] M. A. R. Alif, Yolov11 for vehicle detection: Advancements, performance, and applications in intelligent transportation systems, 2024. doi:10.48550/arXiv.2410.22898.

[30] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, CoRR abs/1405.0312 (2014). URL: http://arxiv.org/abs/1405.0312. arXiv:1405.0312.

[31] A. Torralba, P. Isola, W. T. Freeman, Foundations of Computer Vision, Adaptive Computation and Machine Learning series, MIT Press, 2024. URL: https://mitpress.mit.edu/9780262048972/foundations-of-computer-vision/.