From Plausible to Verifiable: Hybrid Data as a Reference Standard for Explainable and Compliant AI

Julius Schöning[1], Paul Wachter[1]

[1]Osnabrück University of Applied Sciences, Faculty of Engineering and Computer Science, Albrechtstr. 30, 49076 Osnabrück, Germany

Abstract

Trustworthy artificial intelligence (AI) for high-stakes physical systems requires data sufficiency, regulatory compliance, and explanation quality simultaneously. The argument is developed in general terms and instantiated throughout in agricultural robotics, where perception systems must operate under changing lighting conditions, seasonal variation, plant morphology, soil conditions, and in proximity to humans, animals, and machines. This position paper argues that the central challenge lies not only in algorithms but also at the data layer. Hybrid datasets, understood as combinations of real, synthetic, partially synthetic, and augmented samples, provide a structured way to reduce the tension between representative training, legal defensibility, and verifiable explanations. The main contribution is a practitioner-oriented argument that synthetic components in hybrid datasets can serve as a reference standard for explainable AI (XAI). Because the relevant signals, nuisance factors, and controlled artifacts in synthetic samples are known by construction, attribution methods can be evaluated against authored hybrid ground truth rather than judged only by visual plausibility. This hybrid ground truth enables interventional checks in which a factor is inserted, removed, relocated, or decorrelated, and the corresponding prediction and explanation behaviors are measured. The same data-generation process also supports compliance by improving provenance, documentation, auditability, and reproducibility. Hybrid data should therefore be treated as a key system component for trustworthy AI, especially in domains where simulation, controlled generation, and real-world validation can be combined.

Keywords

Hybrid Data, Synthetic Data, Explainable AI, Legal Conform AI, Trustworthy AI, Agricultural Robotics

1. Introduction

Artificial intelligence (AI) systems that perceive and act in the physical world increasingly face requirements that cannot be met solely by predictive performance. In agricultural robotics, the domain that serves as the running example of this paper, perception must remain reliable under changing illumination, occlusion, plant growth stages, soil backgrounds, weather effects, and variable agricultural operational design domains (Ag-ODD) [1, 2]. The same system is also embedded in a regulatory environment that includes data protection, product safety, cybersecurity, data access, and AI regulation [3, 4, 5, 6, 7]. Explanations of model behavior are therefore not an optional interface feature. They are part of debugging, operator trust, certification practice, and post-incident accountability.

The resulting challenge can be described as a three-way tension between data sufficiency, compliance, and explainability. Large real-world datasets improve coverage but may contain personal data, copyrighted content, sensitive telemetry, or unbalanced operational contexts. Legally clean datasets reduce some compliance risk but often lack the long-tail events that matter for safety. Explanation methods can produce convincing visualizations. Nevertheless, their validity is difficult to establish when no independent reference exists for the causal factors used by the AI model [8, 9]. Neither a more complex artificial neural network (ANN) nor a different attribution method removes this structural problem.

This paper argues that hybrid data provide a data-layer response to the tension. Following prior work on hybrid datasets in AI [10, 11], hybrid data are defined here as combinations of real, synthetic, partially synthetic, and augmented samples. The standard motivation for such data is to reduce scarcity. The claim of this position paper is broader: hybrid data can make compliance more intrinsic to the

1st Workshop on Explainability, Transparency, and Safety (ExTraSafe), Bremen, 11. August 2026

j.schoening@hs-osnabrueck.de (J. Schöning); p.wachter@hs-osnabrueck.de (P. Wachter)

0000-0003-4921-5179 (J. Schöning); 0000-0002-6224-6140 (P. Wachter)

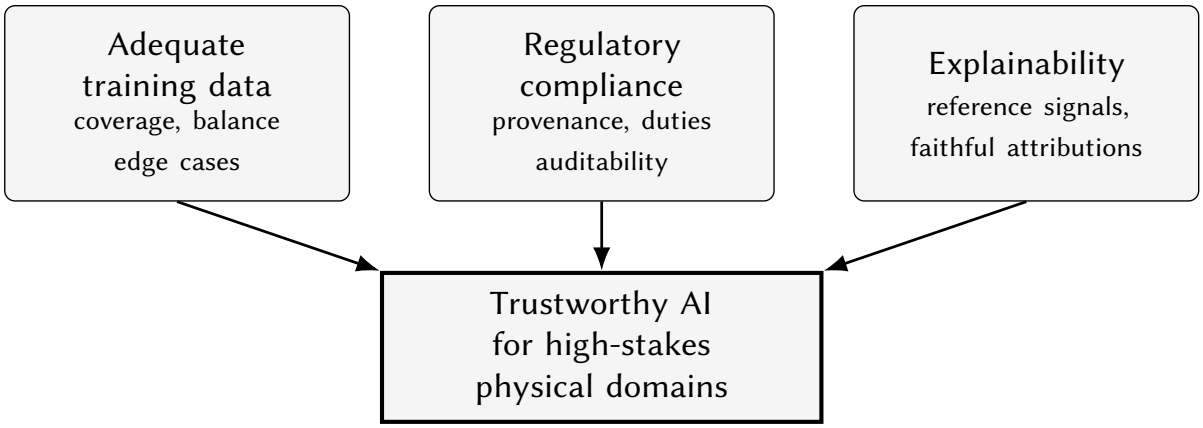

Figure 1: Trustworthy AI in high-stakes domains must meet three demands simultaneously: A data-layer view of trustworthy AI. Hybrid datasets support optimization by expanding coverage, documenting compliance-relevant properties, and providing reference signals for explanation assessment.

data pipeline and can provide a reference standard for evaluating explanations. The contribution is conceptual rather than a new benchmark. It connects hybrid data generation, legal traceability, and explanation evaluation into a single research agenda for trustworthy AI in physical domains. The argument therefore proceeds on two levels: general claims about hybrid data as a component of trustworthy AI, and their concrete instantiation in agricultural robotics, which supplies the examples and the empirical evidence.

Accordingly, the present paper initially characterizes the tension among data, compliance, and explainability. It subsequently clarifies the role of hybrid data and differentiates the legal and explainability functions of controlled generation; these sections develop the general argument, with agricultural robotics providing the concrete instances. The central section elaborates on the assertion that the creation of synthetic content has the potential to transform the evaluation of explanations from visual plausibility to measurement against a known reference, before the limitations and risks of synthetic data are fully recognized. The concluding sections return to the application domain, exploring empirical evidence derived from agricultural hybrid-training benchmarks and extrapolating implications for discussion, dataset reporting, and certification-oriented research.

2. The Data-Compliance-Explainability Tension

Trustworthy AI in high-stakes physical domains must satisfy three requirements simultaneously; agricultural robotics makes each concrete. The data distribution must be broad enough to support robust learning across relevant crop types, growth stages, weather conditions, sensor positions, soil backgrounds, and failure modes. The development process must comply with applicable legal and regulatory obligations [12, 13], including data protection, machine safety, cybersecurity, data access, and AI-specific documentation requirements. The system must also support explanation, because the reasons for a decision may matter to farmers, operators, manufacturers, auditors, and courts. As illustrated in Fig. 1, this requirement is expressed as a constrained optimization problem. The simultaneous fulfillment of thresholds for data adequacy, compliance, and explanation quality is essential for achieving optimal task performance and field fit.

These requirements are individually familiar, but their combination creates a structural tension. A large real-world data corpus may support strong model performance and may be processed under a plausible legal basis, yet it may still leave the model opaque and difficult to audit. A rich dataset that contains workers' faces, geolocated machinery telemetry, or copyrighted images may improve both training and interpretability experiments, but it can create legal exposure under data protection, copyright, or anti-discrimination law. A small dataset that has been lawfully collected and carefully curated may be defensible and easy to inspect, but it will rarely cover rare but safety-relevant agricultural

edge cases.

This tension cannot be resolved completely by AI model selection, training, and optimization. The issue is that the data layer does not merely feed the model. It also defines the observable evidence available for compliance documentation and the reference material available for explanation and evaluation. If the data provide no controlled contrast between signal and background noise, an attribution map can only be inspected for plausibility. If the data provenance is incomplete, later legal review becomes a retrospective reconstruction exercise. A data design that explicitly encodes provenance, variation, and controlled factors is therefore necessary.

3. Hybrid Data as Data-Layer Intervention

Hybrid data combine multiple data sources and generation mechanisms under a defined design. Real data remain essential because they provide a direct connection to the physical world. Synthetic data are generated through simulation, rendering, procedural modeling, or learned generative models. Partially synthetic data combine real and generated content within the same sample, e.g., by placing synthetic weed instances into real field imagery. Augmented data apply label-preserving transformations to existing samples. A hybrid dataset can combine these categories by class, condition, edge case, or training phase, provided that the mixing ratios and generation parameters are recorded.

This definition matters because pure data categories have complementary limitations. Real-world datasets can be costly, legally sensitive, and incomplete for rare scenarios, even when high-quality, as illustrated by agricultural benchmarks such as PhenoBench [14]. Purely synthetic datasets can be perfectly labeled and reproducible; their specific limitations and risks are consolidated in Section 6. Hybrid AI training uses the strengths of each category: realism from measured data, controllability from generated data, and variability from augmentation.

The data-layer intervention has two functions. The first is legal and organizational. Synthetic samples can be generated without identifiable persons, without accidental inclusion of private farm data, and with exact records of the parameters used to create them. The second is epistemic. Synthetic samples encode known causal and nuisance factors, which makes them suitable for testing whether a model and its explanation method respond to the intended content. These functions are connected because the same documentation that supports compliance also supports explanation evaluation.

4. Compliance as Pipeline Property

The compliance function of hybrid data is general, but its stakes become concrete in the application domain. AI-based agrotechnical products fall under several legal categories. They are systems and components of machines that may affect physical safety. Depending on the deployment context, obligations from the General Data Protection Regulation, the Machinery Regulation, the Data Act, and the AI Act may apply concurrently [3, 4, 6, 7]. Prior work on trustworthy AI compliance for farms emphasizes that legal issues are often recognized late and may require expertise beyond the technical development team [15]. A late legal assessment can therefore invalidate design choices, illegal datasets, and prohibited use cases after substantial technical development and research have already been completed.

Hybrid data move part of the legal work upstream into dataset design. Synthetic samples that are not derived from identifiable individuals do not begin as personal data, which changes the compliance posture from after-the-fact anonymization to planned data minimization and provenance control. Partially synthetic pipelines can also specify which content is real, which content is generated, and which transformations were applied. These pipelines improve the ability to document lawful sources, dataset composition, and intended use.

Auditability also changes. A real dataset can be inspected only for the scenarios it contains. A synthetic generator can be queried, parameterized, and rerun to produce coverage of selected scenarios, especially accident or near-accident scenarios in which humans or animals are injured. Questions such

as whether the training data include specific lighting conditions, occlusion types, plant morphologies, soil textures, or sensor artifacts can be answered using generation records rather than relying solely on manual inspection. Reproducibility improves for the same reason: rerunning a generator with recorded seeds and parameters can reproduce the synthetic share exactly, while real-only pipelines can usually provide only an approximation of the original collection conditions.

The assertion advanced here is precise and conducive to discourse: the utilization of hybrid data renders specific compliance-related properties manifest at the data layer. Provenance, intended variation, personal-data avoidance, scenario coverage, and reproducibility can be documented as properties of the pipeline rather than reconstructed at the conclusion of development.

5. Hybrid Data as Reference Standard for Explainable AI

Hybrid Data as Reference Standard for Explainable AI This section develops the domain-independent core of the position, using agricultural robotics as an illustration. Explainability methods for deep visual models often provide post hoc approximations of ANN behavior. Grad-CAM, Integrated Gradients, and LIME are prominent examples that produce attribution maps or local surrogate explanations for opaque predictors [16, 17, 18]. These methods can be useful for debugging and communication, but evaluation remains difficult. A saliency map that highlights a crop region may look reasonable, yet this visual agreement does not prove that the model used the crop for the relevant reason. Shortcut learning is common in modern deep networks, and spurious correlations can yield attributions that appear meaningful while failing under distribution shift [9].

A deeper concern is that current attribution methods provide a spatial answer to a semantic question. A heatmap identifies where features are being read, but it does not certify what concept the highlighted region represents to the model. Rudin [19] argues, from this premise, that post hoc explanations of black-box predictors cannot be relied upon for high-stakes decisions and that inherently interpretable models should be preferred instead. At the same time, Lipton [20] observes that interpretability is itself an underspecified construct, with separate desiderata of trust, causality, transferability, and informativeness that saliency methods address only partially and inconsistently. Empirical work substantiates these critiques. Kindermans et al. [21] demonstrate that several widely used saliency methods fail an input-invariance test, producing different heatmaps for inputs the network treats identically. Tomsett et al. [22] report that the metrics used to evaluate saliency are themselves statistically unreliable, so that the relative ranking of explanation methods depends substantially on the choice of metric. Jacovi and Goldberg [23] crystallise the underlying conceptual error as a confusion of plausibility, an explanation a human finds convincing, with faithfulness, an explanation that accurately reflects model reasoning. A coloured map over crop pixels can be plausible without being faithful, and an agricultural reviewer has no observational means to distinguish the two without an external semantic reference. The field's constructive direction points toward concept-level analysis. Network Dissection [24] labels hidden units with human-interpretable concepts using densely annotated probe data, Concept Bottleneck Models [25] train ANN whose intermediate units predict human concepts directly, and Automatic Concept-based Explanations [26] discover such concepts from segmented super-pixels.

Each of these approaches presupposes the availability of labelled semantic factors, and the impossibility result of Locatello et al. [27] for purely unsupervised disentanglement reinforces this requirement: semantically meaningful factors of variation cannot, in general, be recovered without supervision or strong inductive bias. This is the structural reason hybrid data are valuable for explainability rather than only for training, since the synthetic stratum can carry authored semantic factor masks that real imagery does not provide and cannot easily be made to provide.

The missing element is therefore a reference standard for the validity of explanations with explicit semantic structure. In real-only image data, the relevant causal factors are observed indirectly and often entangled. The presence of a crop, a shadow, a soil texture, a camera artifact, and a field-specific background may be correlated. A human observer can judge whether an attribution map is visually plausible, but this judgment is not equivalent to verification. Sanity checks of saliency maps have shown

that some explanation methods can be insensitive to model parameters or training labels, reinforcing the need for stronger evaluation protocols [8].

Hybrid data can supply such a protocol because synthetic samples are authored. In a generated scene, the plant location, disease symptoms, weed instance, shadow, camera noise, and injected artifact can be recorded during generation. The signal and the background noise factors are known before model inference. A post hoc explanation can therefore be compared with the authored ground truth. If attribution concentrates on the intended biological signal and changes appropriately when that signal is modified, the explanation gains evidential support. If attribution concentrates on an injected watermark, texture, or shadow, the result diagnoses either shortcut learning by the model or a failure of the explanation method. If attribution remains unchanged after the highlighted content is removed, the explanation method itself becomes suspect.

An emerging body of synthetic explainable AI (XAI) benchmarks supports this claim. Yang and Kim [28] formulate the problem directly, noting that no per-feature oracle is available in standard datasets, and propose BAM, a benchmark of pasted foreground objects with authored relative importance, on which several gradient-based attribution methods produce systematic false-positive explanations. Arras et al. [29]. extend this approach with CLEVR-XAI, which provides per-pixel binary ground-truth masks for procedurally rendered visual-question-answering scenes and introduces relevance-mass and relevance-rank accuracy as quantitative metrics for ten widely used attribution methods. Hesse et al. [30] push the same idea to semantic part-level interventions in their FunnyBirds dataset, where individual rendered bird parts can be removed or substituted to evaluate attribution and concept-based explanations against authored part identity. Hooker et al. [31] provide a complementary empirical case for the need for authored ground truth: under their remove-and-retrain protocol, most popular attribution methods perform indistinguishably from random feature assignments on natural-image benchmarks. The concept-level testing tradition of TCAV [32] and the open-source metric library Quantus, which bundles more than thirty evaluation metrics across faithfulness, robustness, and localisation [33], complete the operational infrastructure. Hybrid datasets for agricultural robotics inherit this infrastructure: authored signal masks, background noise masks, and concept labels carried with synthetic samples make existing quantitative XAI metrics directly applicable to plant, soil, and weed classification tasks.

This approach also enables intervention, which is stronger than observation. A generator can produce matched samples in which only one factor changes. The same plant can be rendered with and without a shadow. A spurious marker can be moved away from the crop. A symptom can be suppressed while the background remains constant. The AI model's predictions and explanations can then be evaluated across controlled variants. Such tests separate model dependence on a factor from explanation-method behavior more directly than visual inspection of real images can do. In this sense, hybrid data make explainability experimentally testable.

6. Limitations and Risks of Synthetic Data

The proposed role of synthetic data as a reference standard should not be interpreted as a replacement for real-world validation. The reason is structural rather than methodological. Since the seminal demonstrations of zero-shot transfer by Tobin et al. [34] and Sadeghi and Levine [35], the robotics and computer-vision communities have documented that simulation-trained models routinely exploit rendering artifacts, shading conventions, texture frequencies, or placement regularities that lack real-world counterparts. The Sim2Real workshop summary of Höfer et al. [36] and the survey of Zhao et al. [37] treat this distribution mismatch as a key threat to validity. Domain randomization and related simulation strategies reduce the gap by deliberately varying nuisance parameters during generation [38], and Prakash et al. [39] attribute part of the residual mismatch to context-free placement in unstructured domain randomization, which they partially correct with grammar-aware generation. In the agricultural domain, the risk is concrete: Di Cicco et al. [40] show that crop and weed segmentation networks trained on procedurally rendered fields require careful randomization of soil texture, illumination, and growth stage to approach real-data performance, with characteristic failure modes that lead to

confusion between rendered and real foliage shadows.

For the reference-standard claim of Section 5, the implication is direct. Attributions calibrated on the synthetic stratum of a hybrid dataset cannot be assumed to remain calibrated during deployment, because the model may attend to synthetic-specific cues that do not exist in the real domain, and the explanation method may track them. Two further caveats arise within the synthetic stratum itself. First, controlled artifacts injected for interventional testing can become unrealistic shortcuts that dominate training instead of probing it. Second, authored factor masks certify what was placed in a scene, not what the model internally uses; the evidential weight therefore rests on the interventional protocol, not on the masks alone. The synthetic stratum is consequently a necessary but not sufficient component: faithfulness measurements obtained on it must be cross-validated against the real stratum, and any divergence between the two should be reported as part of the explanation's reliability budget. Controlled synthetic samples create a calibration environment for explanation methods, while real samples test whether the calibrated behavior remains informative under deployment-relevant conditions.

Analogous caveats apply to the compliance argument of Section 4. The claim does not assert that synthetic data inherently ensure legal compliance. When generators are trained on real material, questions of personal data and copyright do not disappear at generation time: generative models can memorize and emit near copies of individual training samples [41], so provenance obligations extend to the generator's training corpus. The documented compliance benefits also presuppose disciplined recording of seeds, parameters, and sources; without such records, the pipeline loses precisely the auditability that motivates its use. Finally, at the training level, growing synthetic shares carry a degradation risk when generated data recursively displace measured data [42, 43]; the real stratum of a hybrid dataset acts as the distributional anchor against this failure mode. None of these limitations argues against hybrid data. They argue for treating the real-to-synthetic composition, the generator provenance, and the synthetic-to-real divergence of explanations as reportable properties of the dataset, as taken up in Section 7.

7. Empirical Basis, Open Questions, and Research Agenda

The conceptual argument depends on hybrid training being viable in practice. If adding synthetic data systematically degraded model performance, the advantages of compliance and explanation would be less useful. Recent benchmarking in agricultural perception provides support for practical viability. Wachter et al. [11] trained 1,800 networks across architectures, hybrid datasets, synthetic-to-real ratios, and mixing strategies; fine-tuning after synthetic pretraining outperformed simple mixed training in 85.2% of paired comparisons. This result does not establish a universal recipe, but it shows that hybrid data can provide a performant foundation when strategies are carefully selected.

For discussion and research, the more salient implication is methodological. Hybrid datasets should be reported not only by size and class distribution, but also by generation mechanism, real-to-synthetic ratio, parameter ranges, seeds, known nuisance factors, and intended edge cases. Explanation studies should report whether an attribution method was evaluated against authored signal masks, nuisance masks, controlled interventions, and matched real samples. Compliance-oriented reports should include provenance, personal-data assumptions, generator inputs, copyright considerations, and scenario coverage. Such reporting would make hybrid datasets comparable across studies and more useful for certification-oriented research.

A concrete research agenda follows from this framing. Dataset designers need protocols for injecting controlled artifacts that avoid the shortcut risk identified in Section 6. Explainability researchers need metrics that compare attribution maps with authored signal and nuisance masks while accounting for spatial uncertainty. Legal and regulatory researchers need documentation templates that translate data-generation records into audit evidence. Agricultural roboticists need deployment studies that compare explanation behavior under synthetic calibration, real validation, and field operation. These problems are well-suited to academic research because they require coordinated discussion across machine learning, AI, simulation, robotics, trustworthiness, explainability, and law.

8. Conclusion

Addressing the tension between data sufficiency, regulatory compliance, and explainability is most effective at the data layer. Hybrid data provides more than additional training samples. They can encode provenance, scenario coverage, reproducibility, and controlled causal variation in a form that supports both technical and legal evaluation. For XAI, the central advantage is the availability of authored ground truth. Synthetic and partially synthetic samples enable testing whether explanations track the intended signal, nuisance variation, or injected artifacts, and controlled interventions make this test stronger than visual plausibility assessment.

Agricultural systems and robotics are a particularly suitable domain for this agenda because they combine known biological mechanisms, simulatable plant and field variation, safety-relevant operations, and significant regulatory stakes. Nevertheless, trustworthy AI and XAI developed for the agricultural domain could serve as a proving ground for methods that may also transfer to industrial automation, environmental monitoring, and medical imaging, and vice versa. The most productive next step is to treat hybrid data as a key component of trustworthy AI systems, alongside reporting standards and evaluation protocols, with equal importance to model architectures and loss functions.

Acknowledgments

This work is part of the AgrifoodTEF-DE project. AgrifoodTEF-DE (reference: 28DZI04A23) is supported by funds of the Federal Ministry of Agriculture, Food and Regional Identity (BMLEH) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the research and innovation program 'Climate Protection in Agriculture'.

Declaration on Generative AI

During the preparation of this work, the authors used the tool Grammarly in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] M. Felske, J. Redenius, G. Happich, J. Schöning, Towards an agricultural operational design domain: A framework, Smart Agricultural Technology 14 (2026) 102246. doi:10.1016/j.atech.2026.102246.

[2] J. Schöning, N. Kruse, Operational design domains and dynamic driving tasks for both off- and on-road agricultural use cases, in: 45. GIL-Jahrestagung, Gesellschaft für Informatik e.V., Bonn, 2025, pp. 179–188. doi:10.18420/GILJT2025_14.

[3] European Parliament and Council of the European Union, Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (GDPR), 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[4] European Parliament and Council of the European Union, Regulation (EU) 2023/1230 of the european parliament and of the council of 14 june 2023 on machinery and repealing directive 2006/42/EC and council directive 73/361/EEC, 2023. URL: https://eur-lex.europa.eu/eli/reg/2023/1230/oj.

[5] European Parliament and Council of the European Union, Directive (EU) 2022/2555 of the european parliament and of the council of 14 december 2022 on measures for a high common level of cybersecurity across the union, amending regulation (EU) no 910/2014 and directive (EU) 2018/1972, and repealing directive (EU) 2016/1148 (NIS 2), 2022. URL: https://eur-lex.europa.eu/eli/dir/2022/2555/oj.

[6] European Parliament and Council of the European Union, Regulation (EU) 2023/2854 of the european parliament and of the council of 13 december 2023 on harmonised rules on fair access to and use of data and amending regulation (EU) 2017/2394 and directive (EU) 2020/1828 (Data Act), 2023. URL: https://eur-lex.europa.eu/eli/reg/2023/2854/oj.

[7] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations and directives (AI Act), 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[8] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

[9] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks, Nature Machine Intelligence 2 (2020) 665–673. doi:10.1038/s42256-020-00257-z.

[10] P. Wachter, N. Kruse, J. Schöning, Synthetic fields, real gains: Enhancing smart agriculture through hybrid datasets, in: 44. GIL-Jahrestagung, Gesellschaft für Informatik e.V., 2024, pp. 437–442. doi:10.18420/GILJT2024_44.

[11] P. Wachter, L. Niehaus, J. Schöning, Development of hybrid artificial intelligence training on real and synthetic data: Benchmark on two mixed training strategies, in: KI 2025: Advances in Artificial Intelligence, Springer Nature Switzerland, 2025, pp. 175–189. doi:10.1007/978-3-032-02813-6_13.

[12] G. Happich, A. Grever, J. Schöning, Agricultural industry initiatives on autonomy: How collaborative initiatives of vdma and aef can facilitate complexity in domain crossing harmonization needs, 2025. doi:10.48550/ARXIV.2501.17962.

[13] J. Schöning, N. Kruse, M. Komesker, D. Barrelmeyer, S. Stiene, Applicable safety tests for automated agricultural machinery and robots: Operational design domains and dynamic driving tasks as key to functional safety, in: LAND.TECHNIK 2024, VDI Verlag, 2024, pp. 341–346. doi:10.51202/9783181024447-341.

[14] J. Weyler, F. Magistri, E. Marks, Y. L. Chong, M. Sodano, G. Roggiolani, N. Chebrolu, C. Stachniss, J. Behley, PhenoBench: A large dataset and benchmarks for semantic image interpretation in the agricultural domain, IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (2024) 9583–9594. doi:10.1109/tpami.2024.3419548.

[15] J. S. Niklas Kruse, Explainable and trustworthy ai compliance for farms, in: 45. GIL-Jahrestagung, Gesellschaft für Informatik e.V., Bonn, 2025, pp. 297–302. doi:10.18420/GILJT2025_32.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 618–626. doi:10.1109/iccv.2017.74.

[17] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3319–3328. URL: https://proceedings.mlr.press/v70/sundararajan17a.html.

[18] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[19] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. doi:10.1038/s42256-019-0048-x.

[20] Z. C. Lipton, The mythos of model interpretability, Communications of the ACM 61 (2018) 36–43. doi:10.1145/3233231.

[21] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, B. Kim, The (un)reliability of saliency methods, in: Explainable AI: Interpreting, Explaining and Vi-

sualizing Deep Learning, Springer International Publishing, 2019, pp. 267–280. doi:10.1007/978-3-030-28954-6_14.

[22] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, A. Preece, Sanity checks for saliency metrics, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, Association for the Advancement of Artificial Intelligence (AAAI), 2020, pp. 6021–6029. doi:10.1609/aaai.v34i04.6064.

[23] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 4198–4205. doi:10.18653/v1/2020.acl-main.386.

[24] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 3319–3327. doi:10.1109/CVPR.2017.354.

[25] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, Concept bottleneck models, in: Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 5338–5348. URL: https://proceedings.mlr.press/v119/koh20a.html.

[26] A. Ghorbani, J. Wexler, J. Zou, B. Kim, Towards automatic concept-based explanations, in: Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf.

[27] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, in: Proceedings of the 36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 4114–4124. URL: https://proceedings.mlr.press/v97/locatello19a.html.

[28] M. Yang, B. Kim, Benchmarking attribution methods with relative feature importance, 2019. doi:10.48550/ARXIV.1907.09701.

[29] L. Arras, A. Osman, W. Samek, CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations, Information Fusion 81 (2022) 14–40. doi:10.1016/j.inffus.2021.11.008.

[30] R. Hesse, S. Schaub-Meyer, S. Roth, FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable ai methods, in: IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2023, pp. 3958–3968. doi:10.1109/iccv51070.2023.00368.

[31] S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim, A benchmark for interpretability methods in deep neural networks, in: Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

[32] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in: Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 2668–2677. URL: https://proceedings.mlr.press/v80/kim18d.html.

[33] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M. M.-C. Höhne, Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond, Journal of Machine Learning Research 24 (2023) 1–11. URL: http://jmlr.org/papers/v24/22-0142.html.

[34] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, Domain randomization for transferring deep neural networks from simulation to the real world, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 23–30. doi:10.1109/iros.2017.8202133.

[35] F. Sadeghi, S. Levine, CAD2RL: Real single-image flight without a single real image, in: Robotics: Science and Systems XIII, 2017. doi:10.15607/rss.2017.xiii.034.

[36] S. Höfer, K. Bekris, A. Handa, J. C. Gamboa, F. Golemo, M. Mozifian, C. Atkeson, D. Fox, K. Goldberg, J. Leonard, C. K. Liu, J. Peters, S. Song, P. Welinder, M. White, Perspectives on Sim2Real transfer for robotics: A summary of the R:SS 2020 workshop, 2020. doi:10.48550/ARXIV.2012.03806.

[37] W. Zhao, J. Peña Queralta, T. Westerlund, Sim-to-Real transfer in deep reinforcement learning for robotics: a survey, in: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2020, pp. 737–744. doi:10.1109/ssci47803.2020.9308468.

[38] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, S. Birchfield, Training deep networks with synthetic data: Bridging the reality gap by domain randomization, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2018, pp. 1082–10828. doi:10.1109/cvprw.2018.00143.

[39] A. Prakash, S. Boochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapira, S. Birchfield, Structured domain randomization: Bridging the reality gap by context-aware synthetic data, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 7249–7255. doi:10.1109/icra.2019.8794443.

[40] M. Di Cicco, C. Potena, G. Grisetti, A. Pretto, Automatic model based dataset generation for fast and accurate crop and weeds detection, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 5188–5195. doi:10.1109/iros.2017.8206408.

[41] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, E. Wallace, Extracting training data from diffusion models, in: 32nd USENIX Security Symposium (USENIX Security 23), USENIX Association, 2023, pp. 5253–5270. URL: https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

[42] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, Y. Gal, AI models collapse when trained on recursively generated data, Nature 631 (2024) 755–759. doi:10.1038/s41586-024-07566-y.

[43] V. Schmitt, N. Kruse, P. Sahitaj, J. Schöning, Transparency as architecture: Structural compliance gaps in eu ai act article 50 ii, in: Proceedings of the Joint Workshop on Legal and Ethical Issues in Human Language Technologies and Computational Approaches to Language Data Pseudonymization, Anonymization, De-identification, and Data Privacy (LEGAL2026 and CALD-pseudo 2026) @ LREC 2026, European Language Resources Association (ELRA), 2026. doi:10.63317/5hwpjzkvb4wf.