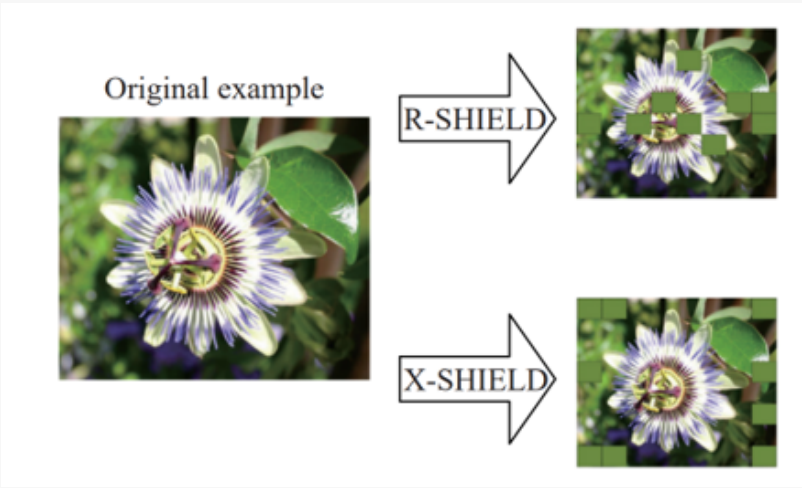
From opaque to accountable: How X-SHIELD Is rewriting the rules of explainable AI

FAYETTEVILLE, GA, UNITED STATES, June 30, 2026 /EINPresswire.com/ -- Artificial intelligence systems are becoming increasingly powerful—but also harder to understand. A new study introduces eXplainable artificial intelligence - SHIELD (X-SHIELD), a regularization technique that improves both the performance and explainability of AI models by selectively hiding the least important features during training.



Transformations proposed by the R-SHIELD and XSHIELD to an example of the Flowers dataset.

For years, the field of eXplainable Artificial Intelligence (XAI) has focused on developing tools to interpret black-box models after they have already been trained—techniques like Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) that generate feature-importance explanations. Yet a critical gap has remained: these explanations were used to understand models, but not to improve them. The idea of closing the loop—using explainability to guide training and enhance model quality—has largely been unexplored. Due to these challenges, there is an urgent need for methods that not only explain AI decisions but also use that explanatory insight to build better, more trustworthy systems.

Now, researchers from the University of Granada's Department of Computer Science and Artificial Intelligence and the Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI) have developed eXplainable artificial intelligence - SHIELD (X-SHIELD), a novel regularization approach published (DOI: 10.1007/s11633-025-1576-y) in the journal Machine Intelligence Research (June 2026). The technique belongs to a broader family called T-SHIELD (Transformation-Selective Hidden Input Evaluation for Learning Dynamics) and represents a concrete step toward what the field calls "Red XAI"—using explainability to improve AI from the developer's perspective.

The core innovation of X-SHIELD lies in how it selects which features to hide. During training, the

model first calculates a saliency map—essentially a gradient-based measure of each feature's importance to the model's decision. The technique then masks out the least important features—for example, background pixels in an image—and computes the Kullback-Leibler divergence between the model's predictions on the original and the modified input. This divergence term is added to the loss function, effectively penalizing the model if its predictions change too much when unimportant features are removed. The result? The model learns to focus on what truly matters. Experiments across seven benchmark image datasets—including CIFAR-10, CIFAR-100, Fashion-MNIST, EMNIST, Flowers, Oxford-IIIT Pet, and ImageNet 1K—showed that X-SHIELD improved accuracy in 13 out of 14 configurations compared to standard training. Perhaps more importantly, the explanations generated by models trained with X-SHIELD were significantly more robust and prescriptive, meaning they better reflected the model's actual decision-making process and remained stable even when the explanation method was run multiple times.

"We realized that explanations were being treated as a final product rather than a tool for improvement," the authors said. "X-SHIELD changes that by making explainability part of the training loop itself. When you force a model to learn without its least important features, it doesn't just become more efficient—it becomes more honest about how it makes decisions. The model can no longer hide behind irrelevant patterns; it has to rely on the features that genuinely matter. And surprisingly, that also makes it more accurate. It's a win-win that we believe could redefine how we think about building trustworthy AI."

The implications extend far beyond academic benchmarks. In high-stakes domains like medical diagnostics, autonomous driving, and financial risk assessment, the ability to trust an AI system is as critical as its raw performance. X-SHIELD offers a practical, plug-and-play solution that can be integrated into existing training pipelines with minimal overhead—roughly a 31% increase in training time for the explainability-guided version, a cost the researchers argue is justified by the gains in transparency and accuracy. Moreover, the method is model-agnostic in the sense that it works with any differentiable architecture, from convolutional neural networks to transformers. As regulations around AI transparency tighten globally, tools like X-SHIELD could become essential for developers seeking to meet both performance benchmarks and accountability standards—making black-box models a little less black, and a lot more reliable.

References
DOI
10.1007/s11633-025-1576-y

Original Source URL
https://doi.org/10.1007/s11633-025-1576-y

Funding Information
This work was supported by the Spanish Ministry of Science and Technology under project (No. PID2023-150070NB-I00) financed by Ministerio de Ciencia e Innovación, Spain (MCIN)/Agencia

Estatal de Investigación (AEI) (Nos. 10.13039 and 501100011033). Funding for open access publishing: Universidad de Granada/CBUA.

Lucy Wang
BioDesign Research
email us here