European Journal of Applied Mathematics (2026), **37**, pp. 1–2
doi:10.1017/S0956792525100235

CAMBRIDGE UNIVERSITY PRESS

EDITORIAL ANNOUNCEMENT

Adversarial robustness of artificial intelligence

Leon Bungert[1], Desmond J. Higham[2] and Laura Thesing[3]

[1]Institute of Mathematics, Center for Artificial Intelligence and Data Science (CAIDAS), University of Würzburg, Würzburg, Germany
[2]School of Mathematics, University of Edinburgh, and Maxwell Institute, Edinburgh, EH3 9FD, UK
[3]Department of Mathematics, Ludwig-Maximilians-Universität München, Munich, Germany
Corresponding author: Desmond J. Higham; Email: d.j.higham@ed.ac.uk

Received: 03 October 2025; **Revised:** 11 October 2025; **Accepted:** 12 October 2025

As systems based on artificial intelligence (AI) and, in particular, deep learning become increasingly prevalent in research, industry, and everyday life, their robustness against adversarial users remains a critical concern. So-called adversarial attacks on image recognition systems or similar classification models have been studied for more than a decade, and the recent advent of transformer technologies and large language models has led to new attack threats with, potentially, more severe consequences. Recent years have seen substantial progress on the algorithmic side of robustifying machine learning, while, at the same time, attacks have become more sophisticated. Indeed, the 2025 *International AI Safety Report* (https://www.gov.uk/government/publications/international-ai-safety-report-2025) from a committee chaired by Yoshua Bengio states that "*Improved understanding of model internals has advanced both adversarial attacks and defences without a clear winner*".

AI is an interdisciplinary field, and a complete understanding of adversarial attacks and defences requires not only extensive empirical evaluations on realistic data sets, but also rigorous theoretical analysis with mathematical tools. In recent years, an increasing number of mathematicians and theoretical computer scientists have identified adversarial robustness in machine learning as a new research field that can be tackled using a range of techniques, including geometric analysis, partial differential equations, optimal transport, applied statistics and numerical analysis.

The purpose of this special issue of the *European Journal of Applied Mathematics* is to grow this community and to provide a high-visibility publication platform for excellent theoretical contributions in the field. The issue contains articles representing both leading experts and early career scientists in the area, spanning a wide range of topics and methods related to adversarial robustness.

Several contributions analyse optimisation problems associated with robustness. García Trillos, Kim and Jacobs consider adversarial training – an optimisation problem that minimises the worst-case loss that an adversary can inflict on a classifier by perturbing inputs. They unify three popular models of multiclass adversarial training and prove the existence of optimal classifiers. The authors also pose the open question of how the results might be extended to parameterised families of neural networks. Frank studies adversarial Bayes classifiers which, by definition, minimise the amount of mislabelling when the data have been attacked by a malicious adversary. The author considers a notion of uniqueness for adversarial Bayes classifiers and connects this to a statistical property of surrogate risks. Morris and Murray answer the question of how fast such adversarial Bayes classifiers converge to the (standard) Bayes classifier as the adversarial budget shrinks. Using geometric arguments, they prove a convergence rate in the Hausdorff distance, thereby improving previously known results in weaker distances.

Ding, Jin, Latz and Liu consider robustness to adversaries that attack using a Bayesian approach, producing a relaxation of the usual minimax problem. They then derive, study and test a computational algorithm for solving the problem. Calder, Drenska and Mosaphir consider a scenario where a player

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0956792525100235 Published online by Cambridge University Press

receives advice from a set of experts about an unknown quantity varying over discrete time. At each step, the player must choose which expert advice to use when making a prediction. The authors derive a continuum limit PDE and an associated numerical method in order to make several conjectures about the optimality of different strategies. Another continuum limit is studied by Weigand, Roith and Burger. They start with the fast signed gradient method (FSGM) – one of the first adversarial attack strategies to be applied to deep learning classifiers. The method may be viewed as computing the most damaging small perturbation, measured in the infinity norm, to an input image. The authors interpret FSGM as an Euler discretization of a differential inclusion related to ∞-curves of maximal slope, allowing several new insights to be drawn.

Robustness verification is the task of guaranteeing that all inputs from a given region are classified correctly. Casadio, Dinkar, Komendantskaya, Arnaboldi, Daggitt, Isac, Katz, Rieser and Lemon consider this issue in the context of natural language processing (NLP). The authors analyse methods to quantify the discrepancy between the verification of the geometric subspaces and the semantic meaning of the sentences that the subspaces represent. They also propose a general NLP verification pipeline.

Liu and Hansen tackle the issue that traditional classification problems have classification functions that are discontinuous across decision boundaries, and hence have an infinite Lipschitz constant. They provide a new stability measure that is shown to provide a more appropriate handle on stability.

Two articles in this special issue focus on fundamental limitations of AI algorithms. Bastounis, Hansen and Vlačić consider deep networks with ReLU activations, showing that for any fixed architecture, there exists a training set for which any network from this class cannot be both accurate and stable. This is despite the existence of a (higher dimensional) network that is accurate and stable on that specially constructed training set. Gazdag, Antun and Hansen look at interval networks, which can be used to quantify the uncertainty associated with a standard neural network. The authors describe a scenario where the task of constructing an optimal pair of interval networks is non-computable.

Cite this article: Bungert L., Higham D. J. and Thesing L. (2026). Adversarial robustness of artificial intelligence. *European Journal of Applied Mathematics*, **37**, 1–2. https://doi.org/10.1017/S0956792525100235

https://doi.org/10.1017/S0956792525100235 Published online by Cambridge University Press