DOI 10.5162/17dss2024/1.1

# Explainable AI – Grundlegende Konzepte und Anwendungen

Nadia Burkart

Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB, Karlsruhe, Deutschland
Kontakt: nadia.burkart@iosb.fraunhofer.de

## Einleitung

In den letzten Jahren hat Künstliche Intelligenz (KI) erhebliche Fortschritte gemacht und sich in zahlreichen Anwendungsbereichen als unverzichtbares Werkzeug etabliert. Insbesondere maschinelle Lernverfahren, wie tiefe neuronale Netze, zeichnen sich durch ihre Fähigkeit aus, aus großen Datenmengen komplexe Muster zu erkennen und präzise Vorhersagen zu treffen. Diese Modelle finden Anwendung in unterschiedlichen Bereichen wie dem Gesundheitswesen, der Finanzindustrie, der industriellen Produktion und dem autonomen Fahren. Trotz ihrer bemerkenswerten Leistungsfähigkeit stoßen diese sogenannten „Black-Box-Modelle“ auf ein zentrales Hindernis: Ihre Entscheidungsfindung bleibt für Menschen weitgehend undurchsichtig. Diese fehlende Nachvollziehbarkeit birgt nicht nur Herausforderungen in der Akzeptanz und im Vertrauen, sondern stellt auch regulatorische und ethische Anforderungen in Frage.

Vor diesem Hintergrund hat das Forschungsfeld der erklärbaren KI (Explainable Artificial Intelligence, XAI) zunehmend an Bedeutung gewonnen. XAI beschäftigt sich mit der Entwicklung von Methoden und Techniken, die darauf abzielen, die Entscheidungsprozesse von KI-Systemen für Menschen verständlich zu machen. Ziel ist es, den „Black-Box“-Charakter der Modelle zu durchbrechen und sie so zu gestalten, dass ihre Vorhersagen und Entscheidungen für Anwender nachvollziehbar und überprüfbar sind. Dies ist besonders in sicherheitskritischen und regulatorisch sensiblen Bereichen wie der Medizin, der Automobilindustrie und der Finanzbranche unerlässlich, wo die Auswirkungen von Entscheidungen weitreichend und potenziell lebensverändernd sein können. Ein zentrales Anliegen von XAI ist die Schaffung von Transparenz, die nicht nur das Vertrauen in KI-Systeme stärkt, sondern auch die Möglichkeit eröffnet, Schwachstellen und Fehlentscheidungen zu identifizieren. Dies ist entscheidend, um Diskriminierung zu vermeiden, ethische Standards einzuhalten und eine robuste Anwendung in der Praxis zu gewährleisten. Darüber hinaus kann die erklärbare KI dazu beitragen, die Effizienz und Präzision von Prozessen zu steigern, indem sie die Zusammenarbeit zwischen Menschen und Maschinen verbessert.

Gleichzeitig steht XAI vor einer Reihe von Herausforderungen. Dazu gehören die Balance zwischen Erklärbarkeit und Modellgenauigkeit, die Komplexität moderner KI-Algorithmen sowie die Entwicklung allgemeingültiger Methoden, die über verschiedene Anwendungsbereiche hinweg einsetzbar sind. Auch der wachsende regulatorische Druck, wie durch den AI-Act, verstärkt die Notwendigkeit, erklärbare Modelle zu entwickeln.

Der AI Act der EU zielt darauf ab, KI-Systeme sicher, transparent und vertrauenswürdig zu gestalten, wobei Erklärbarkeit (Explainability) eine zentrale Rolle spielt. Besonders für Hochrisiko-KI, wie in der Medizin oder Strafverfolgung, schreibt der AI Act vor, dass Entscheidungsprozesse für Nutzer nachvollziehbar gemacht werden müssen. Dies stärkt nicht nur das Vertrauen der Nutzer, sondern ermöglicht auch, Entscheidungen anzufechten und Diskriminierung zu vermeiden. Unternehmen stehen vor der Herausforderung, die Balance zwischen Modellleistung und Erklärbarkeit zu finden, insbesondere bei komplexen Modellen wie neuronalen Netzen. Gleichzeitig soll der AI-Act die Entwicklung neuer XAI-Technologien fördern, die Transparenz und Einhaltung regulatorischer Vorgaben gewährleisten und so einen Wettbewerbsvorteil schaffen können [1] [2]. Erklärbare KI ist somit nicht nur eine technische Notwendigkeit, sondern kann auch ein Schlüssel zur Umsetzung der Anforderungen des AI Act und zur Vertrauensbildung in der europäischen KI-Landschaft sein.

## Grundlegende Konzept im Bereich XAI

Explainable Artificial Intelligence (XAI) umfasst verschiedene Methoden, die darauf abzielen, die Entscheidungsprozesse von KI-Modellen verständlich und transparent zu machen. Eine zentrale Unterscheidung liegt dabei zwischen inhärent interpretierbaren und Black-Box-Modellen. Inhärent interpretierbare Modelle, wie nicht tiefe Entscheidungsbäume oder lineare Regressionsmodelle, bieten von Natur aus Transparenz, da ihre Struktur und Funktionsweise einfach nachvollziehbar sind. Sie sind besonders geeignet für Anwendungen, bei denen Nachvollziehbarkeit im Vordergrund steht, ihre Leistung ist jedoch oft eingeschränkt, wenn komplexe Muster in den Daten vorliegen. Black-Box-Modelle, wie neuronale Netze oder Random Forests, liefern hingegen meist eine höhere Vorhersagegenauigkeit, erfordern jedoch zusätzliche Erklärungsmechanismen, sogenannte post-hoc Methoden, die im Nachgang auf die Modelle draufgesetzt werden, da ihre Entscheidungslogik für den Menschen nicht unmittelbar verständlich ist.

XAI-Methoden werden auch danach unterschieden, ob sie lokale oder globale Erklärungen liefern. Lokale Erklärungen konzentrieren sich auf spezifische Entscheidungen oder Vorhersagen und analysieren, welche Merkmale in einem bestimmten Fall entscheidend waren. Methoden wie SHAP (Shapley Additive Explanations) [3] zu den bekanntesten Ansätzen in

diesem Bereich. Sie quantifizieren den Einfluss einzelner Merkmale und machen Entscheidungen komplexer Black-Box-Modelle lokal verständlich.
Globale Erklärungen hingegen bieten einen Überblick über die generellen Muster und Logiken, die das Modell gelernt hat. Hier kommen Ansätze wie die Analyse der Feature-Wichtigkeit oder die Erstellung von Surrogatmodellen zum Einsatz, die das Verhalten des ursprünglichen Modells approximieren.
Ein weiteres Unterscheidungskriterium ist, ob die Erklärungsmethoden modellagnostisch oder modelleigen sind. Modellagnostische Methoden sind unabhängig vom spezifischen Modelltyp und können auf eine Vielzahl von Black-Box-Modellen angewendet werden. Sie erzeugen Erklärungen durch die Analyse von Eingabe- und Ausgabeparametern und eignen sich daher besonders für Szenarien, in denen die interne Funktionsweise des Modells nicht offengelegt wird. Modellspezifische Methoden hingegen sind speziell auf bestimmte Modelltypen abgestimmt und nutzen deren interne Strukturen. Die Methode Layerwise Relevance Propagation (LRP) [4] erklärt beispielsweise neuronale Netze, indem sie die Bedeutung einzelner Eingaben für die Vorhersage analysiert, während die Feature-Wichtigkeit in Entscheidungsbäumen direkt aus der Modellstruktur abgeleitet wird.
XAI-Methoden adressieren unterschiedliche Dimensionen der Transparenz. Algorithmische Transparenz gibt Einblick in die Funktionsweise des Modells, etwa durch Visualisierungen der internen Prozesse. Ergebnisorientierte Transparenz fokussiert sich auf die Erklärung spezifischer Vorhersagen, indem sie die relevanten Merkmale hervorhebt.
Die Wahl der passenden XAI-Methode hängt stark von der Zielsetzung ab. Erklärungen können zur Fehlerdiagnose in Modellen genutzt werden, um Schwachstellen zu identifizieren, oder zur Einhaltung regulatorischer Anforderungen, insbesondere in sicherheitskritischen Bereichen wie der Medizin oder dem Finanzwesen. Zudem spielt die Benutzerfreundlichkeit eine wichtige Rolle, da Erklärungen sowohl für Experten als auch für weitere Endanwender verständlich sein müssen.
Trotz ihrer Vielseitigkeit stehen XAI-Methoden vor einigen Herausforderungen. Eine der zentralen Schwierigkeiten besteht darin, die Balance zwischen Modellleistung und Erklärbarkeit zu finden. Während einfache Modelle leichter verständlich sind, bieten sie oft nicht die gleiche Leistungsfähigkeit wie komplexe Black-Box-Modelle. Hinzu kommt die Skalierbarkeit, da viele Erklärungsansätze rechenintensiv sind und schwer auf große Datensätze angewendet werden können. Zudem erfordern verschiedene Anwendungsdomänen angepasste Erklärungsansätze, was die Entwicklung universeller Methoden erschwert.
Zusammenfassend bietet die Kategorisierung von XAI-Methoden eine strukturierte Herangehensweise, um passende Erklärungsansätze für spezifische Anwendungen zu identifizieren. Ob lokal oder global, modellagnostisch oder modellspezifisch – die Wahl der Methode hängt stark von den Anforderungen der Domäne und den Einsatzszenarien ab. XAI bleibt ein dynamisches und wachsendes Forschungsfeld, das ständig neue Ansätze entwickelt, um die Erklärbarkeit moderner KI-Systeme zu verbessern.

## Anwendungsbeispiel

Explainable AI (XAI) findet im Bereich des Spritzgießens vielfältige Anwendungsmöglichkeiten. Spritzgießen ist ein verbreitetes Fertigungsverfahren zur Herstellung von Kunststoffteilen, bei dem geschmolzenes Material unter hohem Druck in eine Form eingespritzt wird. Die Qualität der produzierten Teile hängt von einer Vielzahl von Prozessparametern ab, wie Temperatur, Druck, Einspritzgeschwindigkeit und Kühlzeit. Kleine Abweichungen in diesen Parametern können zu Defekten wie Verzug, Lufteinschlüssen oder unvollständigen Teilen führen. Durch den Einsatz von künstlicher Intelligenz und insbesondere von Machine-Learning-Modellen können Muster in den Prozessdaten erkannt und Vorhersagen über die Qualität der produzierten Teile getroffen werden. Sensoren erfassen kontinuierlich Daten wie Temperaturverläufe, Druckprofile und Materialfließverhalten während des Spritzgießprozesses. Diese Daten dienen als Grundlage für Modelle, die Beziehungen zwischen Prozessparametern und Produktqualität herstellen.
Allerdings ist auch hier die fehlende Transparenz der Modelle problematisch. Ein konkreter Anwendungsfall von XAI im Spritzgießen ist die Vorhersage von Defekten in produzierten Teilen. Wenn das Modell prognostiziert, dass ein Teil wahrscheinlich einen Defekt aufweisen wird, können XAI-Methoden eingesetzt werden, um die spezifischen Prozessparameter zu identifizieren, die zu dieser Vorhersage beigetragen haben. Zum Beispiel könnte festgestellt werden, dass ein hoher Einspritzdruck und eine zu kurze Kühlzeit maßgeblich für die erhöhte Defektwahrscheinlichkeit verantwortlich sind, während die Materialtemperatur einen geringeren Einfluss hat.
Ein Verfahren, welches in diesem Bereich angewendet wurde, ist beispielsweise die Permutation Feature Importance. Permutation Feature Importance (PFI) ist eine globale Methode der erklärbaren Künstlichen Intelligenz (XAI), die bewertet, wie wichtig jedes Feature für die Vorhersagen eines Modells ist. Dabei wird gemessen, wie stark die Modellleistung abnimmt, wenn ein bestimmtes Feature permutiert (zufällig durchmischt) wird und somit seine Beziehung zur Zielvariable verloren geht [5]. Die Grundidee besteht darin, die Werte eines Features in den Testdaten zufällig zu permutieren, sodass das Modell keinen Zugriff mehr auf die ursprüngliche Information dieses Features hat. Anschließend wird die Modellleistung, beispielsweise die Genauigkeit oder Fehlerrate, mit den permutierten Daten gemessen und mit der ursprünglichen Leistung verglichen. Eine starke Verschlechterung der Modellleistung weist darauf hin,

dass das Feature wichtig ist, da das Modell ohne dessen Informationen schlechter vorhersagen kann. Bleibt die Leistung unverändert, hat das Feature wenig oder keinen Einfluss auf die Vorhersagen.

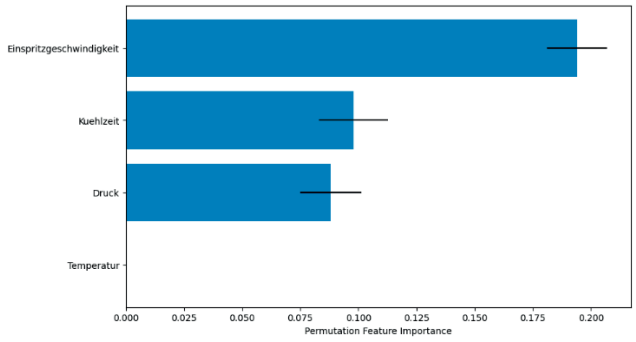


**Abb. 1:** Permutation Feature Importance

Die Einspritzgeschwindigkeit ist auf Abbildung 1 das einflussreichste Feature laut PFI-Analyse, da ihre Permutation zu einer starken Verschlechterung der Modellleistung führt. Sie hat den größten Einfluss auf die Defektwahrscheinlichkeit, da sie stark mit der Zielvariable korreliert ist. Die Kühlzeit ist das zweitwichtigste Feature, ihre Permutation beeinträchtigt die Modellleistung moderat und hat einen spürbaren, aber geringeren Einfluss als die Einspritzgeschwindigkeit. Der Druck hat eine mittlere Bedeutung, da seine Permutation einen geringeren Einfluss auf die Modellleistung hat, jedoch eine gewisse Relevanz zeigt. Die Temperatur ist das einflussärmste Feature, da ihre Permutation keine signifikante Veränderung der Modellleistung verursacht, wodurch sie für ein vereinfachtes Modell möglicherweise entfernt werden könnte.

Ein weiteres Verfahren, welches in diesem Bereich angewendet werden kann, nennt sich Partial Dependence Plots. Ein Partial Dependence Plot (PDP) [5][6]zeigt die Beziehung zwischen einem oder mehreren Features und der Vorhersage eines Modells, während alle anderen Features konstant gehalten werden. Er hilft, den durchschnittlichen Einfluss eines Features oder einer Kombination von Features auf die Modellvorhersage zu verstehen. PDPs bieten eine globale Perspektive, indem sie verdeutlichen, wie sich die Vorhersagewerte ändern, wenn sich der Wert eines Features verändert. Bei zwei Features zeigt ein PDP die Wechselwirkung und den kombinierten Einfluss dieser Features auf die Zielvariable. Diese Methode ist besonders nützlich, um wichtige Features und deren Auswirkungen zu identifizieren, insbesondere bei nicht-linearen Modellen wie Random Forests.

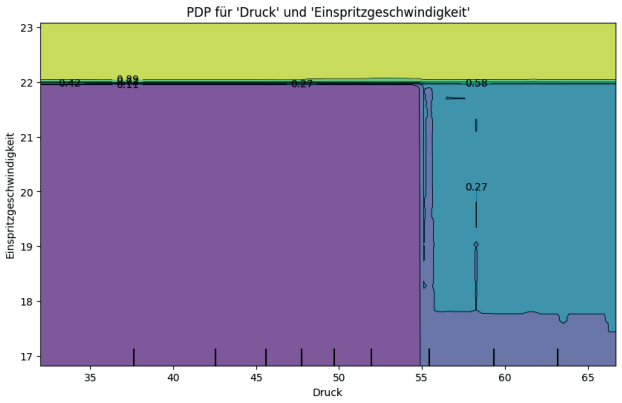


**Abb. 2:** Partial Dependence Plot

Das PDP auf Abbildung 2 zeigt den kombinierten Einfluss der Einspritzgeschwindigkeit (y-Achse) und des Drucks (x-Achse) auf die Vorhersage der Defektwahrscheinlichkeit. Die Farbskala im Diagramm gibt Auskunft über die Defektwahrscheinlichkeit, wobei gelbe Bereiche eine hohe und blaue bis violette Bereiche eine niedrige Wahrscheinlichkeit anzeigen. Einspritzgeschwindigkeiten über 22 führen unabhängig vom Druck zu einer sehr hohen Defektwahrscheinlichkeit, wie der gelbe Bereich oben im Diagramm verdeutlicht. Für Einspritzgeschwindigkeiten unter 22 ist die Defektwahrscheinlichkeit jedoch deutlich geringer, was sich in den blau-violetten Bereichen zeigt. Der Druck hat bei moderaten Einspritzgeschwindigkeiten (unter 22) und Werten unter 55 einen geringen Einfluss auf die Defektwahrscheinlichkeit, sodass diese in diesem Bereich niedrig bleibt (violetter Bereich links unten). Höhere Druck über 55 führen zu einer leichten Erhöhung der Defektwahrscheinlichkeit, wie durch die hellblauen Bereiche sichtbar wird.

Die Interaktion zwischen Druck und Einspritzgeschwindigkeit zeigt, dass hohe Einspritzgeschwindigkeiten den Einfluss des Drucks dominieren. Der Bereich oberhalb einer Einspritzgeschwindigkeit von 22 bleibt gelb, unabhängig davon, wie sich der Druck verändert. Niedrige Einspritzgeschwindigkeiten unter 22 und Druck unter 55 wirken hingegen synergistisch, um die Defektwahrscheinlichkeit zu minimieren. Insgesamt hat die Einspritzgeschwindigkeit den stärksten Einfluss auf die Defektwahrscheinlichkeit, da eine Erhöhung über den Schwellenwert von 22 zu deutlich höheren Wahrscheinlichkeiten führt. Der Druck hat dagegen nur einen mäßigen Einfluss und wirkt primär in Kombination mit der Einspritzgeschwindigkeit. Um die Defektwahrscheinlichkeit zu minimieren, sollte die Einspritzgeschwindigkeit unter 22 und der Druck unter 55 gehalten werden.

Diese Informationen ermöglichen es Ingenieuren, gezielte Anpassungen im Prozess vorzunehmen, wie etwa die Reduzierung des Einspritzdrucks oder die Verlängerung der Kühlzeit, um die Qualität der produzierten Teile zu verbessern. Durch die Visualisierung

der Einflussfaktoren mittels SHAP-Werten können die Zusammenhänge zwischen den Prozessparametern und der Produktqualität besser verstanden werden. Dies fördert nicht nur das Vertrauen in das KI-Modell, sondern liefert auch wertvolle Erkenntnisse zur Prozessoptimierung. Die Anwendung von XAI im Spritzgießen erfordert eine interdisziplinäre Zusammenarbeit zwischen Sensorik, Messtechnik, Informatik und Ingenieurwissenschaften. Sensoren liefern die notwendigen Daten, Messtechniken ermöglichen die genaue Erfassung und Interpretation dieser Daten, und datenbasierte KI-Modelle analysieren sie, um Vorhersagen zu treffen. XAI sorgt schließlich dafür, dass diese Vorhersagen und die zugrunde liegenden Entscheidungsprozesse transparent und verständlich sind. In der Praxis beginnt die Integration von XAI häufig mit Pilotprojekten in kleinem Maßstab, um die Effektivität zu evaluieren. Es ist wichtig, hochwertige und konsistente Daten zu erfassen und die Datenqualität durch Validierungsmechanismen sicherzustellen. Bei erfolgreicher Anwendung kann XAI zu einer erheblichen Reduzierung von Ausschuss, einer Steigerung der Effizienz und einer Verbesserung der Produktqualität führen. Darüber hinaus erleichtert die Erklärbarkeit die Kommunikation zwischen verschiedenen Fachbereichen und fördert die Akzeptanz von KI-Technologien in der Produktion.
Zukünftig könnten KI-Systeme mit XAI-Funktionen noch stärker in den Produktionsprozess integriert werden, indem sie nicht nur Vorhersagen und Erklärungen liefern, sondern auch automatische Anpassungen an den Maschinen vornehmen. Die Kombination von Sensorik, Messtechnik und erklärbarer KI im Spritzgießen zeigt, wie Innovationen durch die Vernetzung verschiedener Disziplinen entstehen.

## Zusammenfassung und Ausblick

Künstliche Intelligenz (KI) hat sich in den letzten Jahren als unverzichtbares Werkzeug in verschiedenen Branchen etabliert. Besonders maschinelle Lernverfahren, wie tiefe neuronale Netze, haben durch ihre Fähigkeit, komplexe Muster zu erkennen, zahlreiche Anwendungen in Bereichen wie Gesundheitswesen, Finanzindustrie, Produktion und autonomem Fahren gefunden. Jedoch stoßen diese „Black-Box-Modelle“ auf Herausforderungen hinsichtlich Nachvollziehbarkeit und Akzeptanz. Das Forschungsfeld der erklärbaren KI (Explainable AI, XAI) hat daher das Ziel, die Entscheidungsprozesse moderner KI-Modelle verständlich und transparent zu machen. XAI trägt dazu bei, Vertrauen zu schaffen, Schwachstellen zu identifizieren und regulatorische Anforderungen zu erfüllen, wie sie etwa durch den EU AI Act gestellt werden. Grundlegende Konzepte in XAI unterscheiden zwischen lokalen und globalen Erklärungen sowie zwischen modellagnostischen und modellspezifischen Methoden. Beispiele wie SHAP oder LIME quantifizieren den Einfluss einzelner Merkmale auf Entscheidungen, während Surrogatmodelle globale Muster darstellen. Die Anwendung von XAI reicht von der Fehleranalyse über die Einhaltung regulatorischer Vorgaben bis hin zur Verbesserung der Mensch-Maschine-Kollaboration. Trotz ihrer Vielseitigkeit steht XAI vor Herausforderungen wie der Balance zwischen Erklärbarkeit und Modellleistung oder der Entwicklung domänenspezifischer Lösungen.
Die Bedeutung von XAI wird in den kommenden Jahren weiter zunehmen, insbesondere im Kontext regulatorischer Anforderungen wie dem EU AI Act und der wachsenden Akzeptanz von KI-Systemen in sicherheitskritischen Bereichen. Die Forschung wird sich darauf konzentrieren, weiter neue Methoden zu entwickeln, die universell einsetzbar sind, ohne dabei die Genauigkeit der Modelle zu beeinträchtigen. Gleichzeitig wird die Interdisziplinarität eine zentrale Rolle spielen, da die Integration von Sensorik, Messtechnik und KI-Technologien neue Potenziale in Produktion und anderen Anwendungsfeldern eröffnet. Langfristig werden erklärbare KI-Systeme eine Schlüsselrolle in der Entwicklung von menschenzentrierten Technologien spielen, indem sie Vertrauen und Transparenz in komplexe technologische Prozesse bringen. Dies wird nicht nur die industrielle Praxis transformieren, sondern auch die gesellschaftliche Akzeptanz von KI nachhaltig stärken.

## Literatur

[1] Panigutti, Cecilia, et al. "The role of explainable AI in the context of the AI Act." *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 2023.

[2] Burkart, Nadia, and Marco F. Huber. "A survey on the explainability of supervised machine learning." *Journal of Artificial Intelligence Research* 70 (2021): 245-317.

[3] Chen, Hugh, Scott M. Lundberg, and Su-In Lee. "Explaining a series of models by propagating Shapley values." *Nature communications* 13.1 (2022): 4512.

[4] Montavon, Grégoire, et al. "Layer-wise relevance propagation: an overview." *Explainable AI: interpreting, explaining and visualizing deep learning* (2019): 193-209.

[5] Molnar, Christoph. *Interpretable machine learning*. Lulu. com, 2020.

[6] Casalicchio, Giuseppe, Christoph Molnar, and Bernd Bischl. "Visualizing the feature importance for black box models." *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer International Publishing, 2019.