Paul P. Martin[1]
Amber J. Dood[2]
Field M. Watts[3]
Ginger V. Shultz[2]
Nicole Graulich[1]

[1]Justus-Liebig-Universität Gießen
[2]University of Michigan
[3]University of Wisconsin-Milwaukee

Wie vertrauenswürdig ist maschinelles Lernen in der Chemie?

Lückenhafte Richtlinien zur Prüfung der Vertrauenswürdigkeit von ML-Modellen

Die Anwendung maschinellen Lernens (ML) nimmt in der naturwissenschaftsdidaktischen Forschung und Praxis stetig zu (z. B. Lee, Yun, Zhai & Krippen, 2025). Mit der zunehmenden Anwendung von ML in bildungsbezogenen Kontexten wachsen jedoch auch Bedenken hinsichtlich der Vertrauenswürdigkeit solcher Modelle (z. B. Grimm, Steegh, Kubsch & Neumann, 2023). Diese Bedenken betreffen insbesondere drei Aspekte: die Gestaltung und Validität der zugrunde liegenden Assessments, die technischen Eigenschaften und Grenzen der ML-Modelle sowie potenzielle Verzerrungen in der automatisierten Analyse von Lernenden-Daten (z. B. Cheuk, 2021; Pellegrino, DiBello & Goldmann, 2016; Rupp, 2018).

Um diesen Herausforderungen zu begegnen, wurden in Disziplinen wie der Informatik oder der Bildungspsychologie bereits verschiedene Konzepte und Richtlinien zur Validierung von ML-Modellen vorgeschlagen (z. B. Kapoor, Cantrell, … & Narayanan, 2024; Rupp, 2018; Williamson, Xi & Breyer, 2012). Diese Ansätze liefern wertvolle Impulse, zeigen jedoch in der Praxis eine stark heterogene Umsetzung (Martin & Graulich, 2023; Zhai, Yin, Pellegrino, Haudek & Shi, 2020). Während einige Studien lediglich die Übereinstimmung der maschinellen Klassifikation mit menschlicher Kodierung berichten, verwenden andere aufwendigere Validierungsverfahren, etwa durch Mehrfachvalidierung, externe Testdatensätze oder erklärbare KI-Methoden. Insgesamt mangelt es an einheitlichen Richtlinien zur Entwicklung, Validierung und Berichterstattung von ML-Modellen in der naturwissenschaftsdidaktischen For--chung.

Darüber hinaus ist eine bloße Übernahme bestehender Richtlinien für die Naturwissenschafts-didaktik nicht ausreichend. Es stellt sich nämlich die Herausforderung, dass viele der potenziellen Nutzergruppen – darunter Lehrpersonen, Lernende, Eltern oder Akteure der Bildungs--olitik – keine Experten im Bereich der Anwendung von ML-Modellen sind. Richtlinien müs--en daher so gestaltet sein, dass sie nicht nur eine technische Validierung, sondern auch eine transparente Kommunikation der Ergebnisse, Anwendungsbereiche, Stärken und Limitationen von ML-Modellen ermöglichen. Hier bietet das Konzept der Vertrauenswürdigkeit (*engl.* trustworthiness) einen vielversprechenden Ansatz (z. B. Varshney, 2022). Es zielt darauf ab, Anwendern eine evidenzbasierte Grundlage zu liefern, um ein angemessenes Maß an Vertrauen in ML-Modelle zu entwickeln.

Einführung von Richtlinien zur Vertrauenswürdigkeit von ML-Modellen

Um die zuvor beschriebenen Herausforderungen beim Einsatz von ML in der Bildungsforschung systematisch zu adressieren, bedarf es einheitlicher Richtlinien. Dabei bietet es sich an, auf bewährte Kriterien aus der qualitativen Bildungsforschung zurückzugreifen, namentlich auf die sogenannten Gütekriterien der *Trustworthy Naturalistic Inquiry* (Guba, 1981). Diese Kriterien – *Credibility*, *Dependability*, *Transferability* und *Confirmability* – sind auf

eine möglichst umfassende und nachvollziehbare Prüfung der Vertrauenswürdigkeit qualitativer Forschungsergebnisse ausgerichtet. Ihre Übertragung auf die ML-Methodik schafft eine geläufige Grundlage, um die Vertrauenswürdigkeit von ML-Modellen in der fachdidaktischen Forschung greifbar, kommunizierbar und beurteilbar zu machen.

Im Rahmen dieser Übertragung lassen sich vier zentrale Gütekriterien des *Trustworthy Machine Learning* identifizieren, die analog zu den klassischen Kriterien strukturiert sind: *Performance*, *Reliability*, *Aligned Purpose* und *Human Interaction* (Varshney, 2022). Jedes dieser Gütekriterien steht in direkter Verbindung zu einem der etablierten Kriterien vertrauenswürdiger Forschung (Abb. 1):

1. *Performance* entspricht dem Kriterium der *Credibility*. Während *Credibility* in der qualitativen Forschung die Frage behandelt, inwieweit Forschungsergebnisse die Perspektiven der Teilnehmenden authentisch widerspiegeln, zielt das Gütekriterium *Performance* im Kontext von ML auf die Genauigkeit der Vorhersagen ab. Es beschreibt, wie zuverlässig ein ML-Modell seine beabsichtigte Aufgabe erfüllt – etwa die automatisierte Bewertung von Lernendenantworten – und wie hoch die Übereinstimmung zwischen maschinellen Vorhersagen und menschlicher Kodierung ist. Beide Kriterien teilen also das Interesse an der Genauigkeit der jeweiligen Analyse.
2. *Reliability* steht im engen Zusammenhang mit dem Kriterium der *Dependability*. Während *Dependability* die Replizierbarkeit eines Forschungsprozesses betont, meint *Reliability* im ML-Kontext die Stabilität und Übertragbarkeit eines ML-Modells auf andere Aufgaben, Datensätze oder Kontexte. Ein zuverlässiges ML-Modell erzielt auch unter veränderten Bedingungen vergleichbare Ergebnisse. Dieser Anspruch verbindet die Prinzipien systematischer Forschung mit den Anforderungen an robuste ML-Modelle.
3. *Aligned Purpose* lässt sich auf das Kriterium der *Transferability* übertragen. In der qualitativen Forschung beschreibt *Transferability* die Frage, inwieweit sich Forschungsergebnisse auf andere Kontexte übertragen lassen, was voraussetzt, dass der ursprüngliche Kontext ausreichend detailliert beschrieben ist. Analog dazu beschreibt *Aligned Purpose*, ob das ML-Modell zweckgemäß eingesetzt wird und ob sein Design, seine Trainingsdaten und seine Anwendung ethisch, didaktisch und funktional auf das Ziel abgestimmt sind, das es erfüllen soll. Die Aussagekraft eines ML-Modells hängt somit davon ab, ob es für den vorgesehenen Zweck angemessen entwickelt und eingesetzt wird.
4. *Human Interaction* hängt schließlich mit dem Kriterium der *Confirmability* zusammen. *Confirmability* fordert die Transparenz der Interpretation von Forschungsergebnissen, etwa durch Reflexion möglicher Verzerrungen oder offengelegte Analysepfade. Im ML-Kontext meint *Human Interaction* die Frage, ob die Entscheidungsprozesse eines Modells für Menschen verständlich sind, etwa durch die Nutzung interpretierbarer Modellarchitekturen oder erklärbarer KI-Techniken (z. B. Local Interpretable Model-agnostic Explanations (LIME) oder SHapley Additive exPlanations (SHAP)). Interpretierbarkeit ist im Bildungsbereich essenziell, um Lehrpersonen, Lernenden oder Entscheidungsträgern eine reflektierte Auseinandersetzung mit dem Modelloutput zu ermöglichen.

Anwendung der Richtlinien zur Vertrauenswürdigkeit von ML-Modellen

Im Rahmen einer empirischen Fallstudie in der Hochschullehre der Organischen Chemie wurden die postulierten Gütekriterien angewendet, um die Vertrauenswürdigkeit eines ML-Modells zu prüfen (Martin, Dood, Watts, Shultz & Graulich, under review). Das ML-Modell wurde entwickelt, um die Qualität von Studierendenantworten zum mechanistischen Begründen automatisch zu bewerten. Studierende der Organischen sollten dazu in kurzen Texten be-

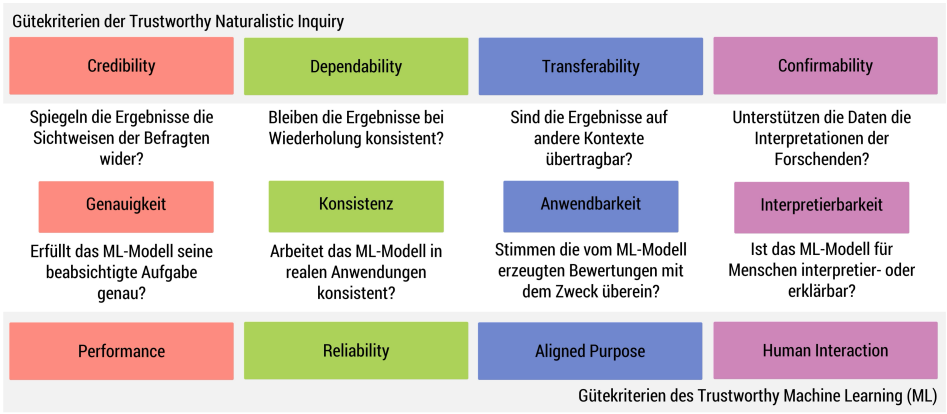

Abb. 1: Gemeinsame Gütekriterien der Trustworthy Naturalistic Inquiry und des Trustworthy Machine Learning

gründen, ob ein gegebenes Reaktionsprodukt plausibel ist (Lieber & Graulich, 2022; Lieber, Ibraj, Caspari-Gnann & Graulich, 2022). Zur automatisierten Bewertung wurde ein supervised ML-Modell entwickelt, das auf fortgeschrittenen Deep-Learning-Architekturen basiert (Martin, Kranz, Wulff & Graulich, 2024; Martin & Graulich, 2024a, 2024b). Das Modell nutzt ein vortrainiertes transformerbasiertes Sprachmodell, konkret *BERT-base-uncased* (Devlin, Chang, Lee & Toutanova, 2019).

Die Analyse des ML-Modells anhand organisch-chemischer Lernaufgaben zeigte, dass die vier untersuchten Gütekriterien – *Performance*, *Reliability*, *Aligned Purpose* und *Human Interaction* – zwar jeweils eigenständige Prüfgrößen darstellen, jedoch in einem komplexen Spannungsverhältnis zueinanderstehen. So konnte beispielsweise gezeigt werden, dass ML-Modelle, die auf spezifische Kontexte zugeschnitten waren, eine besonders hohe Genauigkeit in diesen Kontexten aufwiesen. Diese Genauigkeit ließ jedoch deutlich nach, sobald dieselben Modelle in anderen Aufgabenformaten oder an weiteren Institutionen eingesetzt wurden, was auf eine eingeschränkte Konsistenz und Anwendbarkeit hinweist. Umgekehrt verbesserte sich die Konsistenz über Kontexte hinweg, wenn Trainingsdaten aus mehreren Kontexten kombiniert wurden; dies ging allerdings zulasten der ursprünglichen Modellgenauigkeit. Die Gütekriterien Genauigkeit und Konsistenz stehen somit in einem potenziellen Zielkonflikt.

Ein weiteres Spannungsfeld zeigte sich zwischen Genauigkeit und Interpretierbarkeit: Während komplexere Modelle (z. B. neuronale Netze) hohe Genauigkeit erzielten, erschwerte ihre Komplexität die Interpretierbarkeit der getroffenen Entscheidungen für menschliche Nutzer. Vereinfachte, interpretierbare Modelle boten zwar mehr Transparenz, konnten jedoch fachspezifische Begründungsmuster weniger zuverlässig abbilden.

Insgesamt zeigen die Ergebnisse, dass die Vertrauenswürdigkeit von ML-Modellen nicht durch ein einzelnes Gütekriterium, sondern durch das ausgewogene Zusammenspiel mehrerer, teils konkurrierender Gütekriterien bestimmt wird. Die bewusste Reflexion dieser Spannungsverhältnisse ist zentral, um fundiertes Vertrauen in ML-Anwendungen in den Naturwissenschaftsdidaktiken zu ermöglichen. Die Anwendbarkeit ist hier ein zentrales vermittelndes Gü--ekriterium: Die Passung eines ML-Modells zu seinem konkreten Einsatzzweck ist entschei--end dafür, inwiefern die Gütekriterien zu gewichten sind und ob Abstriche in dem einen oder anderen Gütekriterium gerechtfertigt erscheinen.

Literatur

Cheuk, T. (2021). Can AI be racist? Color-evasiveness in the application of machine learning to science assessments. *Science Education, 105*(5), 825-836.

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Association for Computational Linguistics: Human Language Technologies*, Volume 1 (pp. 4171-4186)

Grimm, A., Steegh, A., Kubsch, M. & Neumann, K. (2023). Learning analytics in physics education: Equity-focused decision-making lacks guidance! *Journal of Learning Analytics, 10*(1), 71-84

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology Journal, 29*(2), 75-91.

Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., Hofman, J. M., Hullman, J., Lones, M. A., Malik, M. M., Nanayakkara, P., Poldrack, R. A., Raji, I. D., Roberts, M., Salganik, M. J., Serra-Garcia, M., Stewart, B. M., Vandewiele, G. & Narayanan, A. (2024). Reforms: Consensus-based recommendations for machine-learning-based science. *Science Advances, 10*(18), Article eadk3452, 1-17 .

Lee, G., Yun, M., Zhai, X. & Crippen, K. (2025). Artificial intelligence in science education research: Current states and challenges. *Journal of Science Education and Technology*, Early View, 1-18.

Lieber, L. S. & Graulich, N. (2022). Investigating students' argumentation when judging the plausibility of alternative reaction pathways in organic chemistry. *Chemistry Education Research and Practice, 23*(1), 38-53.

Lieber, L. S., Ibraj, K., Caspari-Gnann, I. & Graulich, N. (2022). Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns. *Chemistry Education Research and Practice, 23*(4), 811-828.

Martin, P. P., Dood, A. J., Watts, F. M., Shultz G. V. & Graulich, N. (under review). To what extent can we trust a machine? Establishing trustworthiness guidelines to evaluate supervised machine learning models across chemistry contexts. *International Journal of Artificial Intelligence in Education*.

Martin, P. P. & Graulich, N. (2023). When a machine detects student reasoning: A review of machine learning-based formative assessment of mechanistic reasoning. *Chemistry Education Research and Practice, 24*(2), 407-427.

Martin, P. P. & Graulich, N. (2024a). Beyond language barriers: Allowing multiple languages in postsecondary chemistry classes through multilingual machine learning. *Journal of Science Education and Technology, 33*(3), 333-348.

Martin, P. P. & Graulich, N. (2024b). Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence. *Computers and Education: Artificial Intelligence, 7*, Article 100265, 1-15.

Martin, P. P., Kranz, D., Wulff, P. & Graulich, N. (2024). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching, 61*(8), 1757-1792.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59-81.

Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education, 31*(3), 191-214.

Vashney, K. R. (2022). *Trustworthy machine learning*. Independently published.

Williamson, D. M., Xi, X. & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C. & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education, 56*(1), 111-151.