WIAS-Richtlinie zur Nutzung generativer KI

Präambel

Der Einsatz von *generativer künstlicher Intelligenz* (**GenKI**) hat sich in der Gesellschaft und der Wissenschaft weitestgehend verselbständigt. Es besteht ein Konsens, dass diese disruptive Technologie ein großes positives Potenzial für Lehre und Forschung besitzt, welches am WIAS ausgeschöpft, gesteuert und begleitet werden soll. Das Ziel der wissenschaftlichen GenKI-Nutzung muss ein Erkenntnisgewinn sein, der das Vertrauen der Gesellschaft in wissenschaftliche Methoden und deren Ergebnisse stärkt und zugleich die Freiheit von Forschung und Lehre wahrt. Für die Nutzung generativer KI am WIAS sind maßgeblich:

- Europäische *KI-Verordnung* EU 2024/1689 [EU AI Act],
- Europäische *Datenschutz-Grundverordnung* [DSGVO],
- *Leitlinien Guter Wissenschaftlicher Praxis* von DFG und Forschungsverbund Berlin e.V. (FVB) [GWP],
- *Living guidelines on the responsible use of generative AI in research* der Europäischen Kommission [EU Guideline],
- Empfehlung zur Sicherung der guten wissenschaftlichen Praxis beim Umgang mit Künstlicher Intelligenz der Leibniz Gemeinschaft [Leibniz Guideline],
- und ggf. die Richtlinien von Journalen, Kooperationspartnern und Universitäten.

Diese Richtlinie richtet sich an Beschäftigte des WIAS in Forschung, Verwaltung und IT und ebenso an WIAS-Gäste, die GenKI-Systeme des WIAS nutzen. Sie regelt primär die dienstliche Nutzung sogenannter *Large Language Modelle* **(LLM)**. Die Kernprinzipien der Forschung mit und an KI sind [EU Guideline]:

- **Verlässlichkeit:** Forschungsqualität durch Überprüfbarkeit und Reproduzierbarkeit sichern und auf Verzerrung (Bias) achten.
- **Redlichkeit:** Forschung transparent, fair und unparteiisch; Einsatz generativer KI offenlegen.
- **Respekt:** Achtung von Mensch, Gesellschaft, Umwelt, sensiblen Daten (z.B. personenbezogene Daten, Geschäftsgeheimnisse, Gutachten, Anträge, geistig geschützte Inhalte), Privatsphäre, geistigen Eigentum und Fairness.
- **Verantwortlichkeit:** Forschende bleiben für alle durch KI generierten Ergebnisse und deren Folgen verantwortlich; Nutzung generativer KI erfolgt unter menschlicher Aufsicht und Steuerung.

Verbotene KI-Praktiken im Sinne des [EU AI Act] sind am WIAS auch in der Forschung ausgeschlossen; Forschung und Entwicklung am WIAS richten sich nicht auf solche Zwecke.

Das WIAS kommt mit dieser Richtlinie und ihren begleitenden Prozessen zugleich den Pflichten aus Art. 4 und Art. 11 KI-VO nach, ein angemessenes Maß an KI-Kompetenz sicherzustellen und eine technische Dokumentation für die Nutzung der angebotenen KI-Systeme zur Verfügung zu stellen. Die Gesamtverantwortung liegt bei der Institutsleitung und die Umsetzung erfolgt durch die in dieser Richtlinie genannten Rollen (u.a. Datenschutzkoordination, Ombudsperson, RSE, IT). Weiterbildungen zu KI-Themen sollen in der Zukunft angeboten werden; z.B. im Rahmen des RSE-Seminars am WIAS gibt es Veranstaltungen, um die KI-Kompetenz zu gewährleisten.

Stufensystem (TIER 0,1,2)

Im Sinne des [EU AI Act] ist *wissenschaftliche Forschung* in der Regel mit einem *niedrigen Risiko* verbunden. Zur weiteren praktischen Abschätzung und Eingrenzung von Risiken bei der Nutzung eines LLM setzt das WIAS ein TIER-Stufensystem ein. Es unterscheidet Art und Zweck des LLM, ersetzt aber nicht die offizielle Hochrisiko-/Niedrigrisiko-Einstufung (und verbotene Praktiken) des [EU AI Act]. Die wesentlichen Merkmale der TIER-Stufen sind in Tabelle 1 abgebildet. Forschende können je nach Komplexität der Forschungsfragen TIER 0, TIER 1 oder TIER 2 LLM nutzen. Die Entwicklung, Anpassung und Evaluation eigener KI-Modelle ist im Rahmen der TIER-Stufen ausdrücklich möglich. Dabei werden die maßgeblichen Rechtsvorschriften [DSGVO, EU AI Act] sowie die einschlägigen Leitlinien [GWP, EU Guideline, Leibniz Guideline] eingehalten. [DSGVO] und [EU AI Act] definieren bestimmte Ausnahmen und Privilegien für die Forschung (z.B. Art. 89 [DSGVO] oder Erwägungsgrund 25 [EU AI Act]).

TIER	Risiko	Typische Modelle / Nutzung	Anforderungen
0	kaum	lokale LLM ohne externe Datenübertragung; Basistools zur Text- und Codeunterstützung	automatischer Zugang; Beachtung von Guideline und [GWP, DSGVO]
1	niedrig	kuratierte, externe LLM mit Standardfunktionen (Chat, Text, Code) für Forschung und Administration ohne sensible Daten	Kenntnisnahme und Einhaltung der Richtlinie (**schwarz**) und [GWP, DSGVO]; Beschränkung auf nicht-sensible Daten
2	mittel	leistungsfähige externe LLM, spezialisierte APIs, komplexe agentische Workflows; Forschung *mit* und *an* Modellen	formeller Antrag + Prüfung, unterschriebene Richtlinie TIER 2 (**blau**), erweiterte Anforderungen und technische Expertise

Tabelle 1: Stufensystem für KI-Nutzungsfälle am WIAS

Die Nutzung von LLM in der Verwaltung unterliegt den rechtlichen Vorgaben von [DSGVO, EU AI Act]. Insbesondere Vorgänge in der Personal- und Finanzverwaltung, der Drittmittel- und Vertragsverwaltung sowie der internen Organisation unterliegen einer besonderen *Sorgfaltspflicht*. Für die Verwaltung sollen daher in der Regel nur TIER 0 und nur bei klar unbedenklichen[1] Daten TIER 1 LLM genutzt werden. KI-Systeme dürfen nicht zur automatisierten Bewertung von Bewerbenden oder Beschäftigten benutzt werden.

Für TIER 1 und TIER 2 sind jeweils auch Token- und Kostenlimits der externen LLM-Anbieter zu beachten: die Modelle sind entsprechend differenziert nach Anwendungszweck energie- und kostensparend einzusetzen. In den Stufen TIER 0 und TIER 1 erfolgt die Nutzung ausschließlich aus dem WIAS-Intranet/VPN über

`https://webui.wias-berlin.de`.

In der Stufe TIER 2 erfolgt der Zugriff auf externe LLM ggf. direkt und somit nicht notwendig aus dem WIAS-Intranet/VPN. Der Freigabeprozess von TIER 2 erfolgt durch einen **Antrag** (siehe Tabelle 1).

Richtlinie

Forschende am WIAS dürfen generative künstliche Intelligenz wie andere Werkzeuge zur Unterstützung in Forschung, Lehre, Softwareentwicklung und Verwaltung einsetzen, sofern sie die Regeln der [GWP], [DSGVO], der Universitäten und der Journale beachten. Die Verantwortung für die Richtigkeit und Verwendbarkeit der Inhalte (Input und Output) verbleibt immer beim KI-Nutzenden. Für alle WIAS-Beschäftigten sind bei KI-Nutzung geltende Rechtsbestimmung zu beachten, z.B. hinsichtlich des Datenschutzes, Urheberrechts, Patentrechts.

Im Folgenden werden zentrale Aspekte bei der Nutzung von GenKI hervorgehoben.

Verhaltensregeln und Dokumentation in der Forschung

1. Die Leitlinien guter wissenschaftlicher Praxis des Forschungsverbunds Berlin e.V. und der DFG [GWP] sind Grundlage allen wissenschaftlichen Arbeitens am WIAS. GenKI darf die eigenständige wissenschaftliche Leistung nicht ersetzen, sondern nur unterstützen. Ausgaben sind auf inhaltliche Richtigkeit (Halluzinationen) und Bias zu prüfen.

2. Die Nutzung von GenKI ist für wissenschaftliche Arbeiten (Publikationen, Gutachten, Berichte, Lehrmaterialien etc.) so zu dokumentieren, dass Art, Umfang und Zweck der Unterstützung bzw. die eigene wissenschaftliche Leistung nachvollziehbar sind.

3. Autorenschaft, Verantwortlichkeit und Haftung verbleiben bei den jeweiligen Nutzenden und ggf. Co-Autor*innen. GenKI kann nicht als Autor*in ausgewiesen werden.

TIER 2: Um das Risiko der Nutzung nachvollziehbar zu gestalten, ist ein der Komplexität der Daten angemessener Forschungsdatenmanagementplan (RDMP) anzufertigen. Bei einer Änderung des Datenprofils wird der RDMP angepasst. In einfachen Fällen dient der Antrag als RDMP. Besonders im TIER 2 ist eine hohe eigene KI-Kompetenz erforderlich und die Nutzenden verpflichten sich, ihre Kenntnisse zum Stand von Technik und Wissenschaft aktuell zu halten.

Datenschutz Beim Einsatz von KI-Systemen gelten uneingeschränkt dieselben datenschutzrechtlichen Grundsätze und Anforderungen wie bei allen anderen Anwendungen und Prozessen. Für den datenschutzkonformen Einsatz von KI-Systemen sind insbesondere die folgenden Punkte zu beachten:

4. Der Nutzende ist verantwortlich für die Rechtmäßigkeit der Eingaben in und der Verwertung von

[1] Klar unbedenkliche Daten erhalten keine persönliche, urheberrechtlich geschütze oder vertrauliche Informationen.

Ausgaben von LLM. Externe LLM (ab TIER 1) können Ein- und Ausgaben für einige Zeit speichern und automatisiert überwachen, um Rechtsverstöße zu prüfen. Nach derzeitigem Kenntnisstand werden Eingaben über die API an externe LLM aktuell nicht zum Training verwendet. Feedbacks (z.B. Daumen hoch/runter) können jedoch gesondert geregelt sein und unter Umständen dennoch zum Training benutzt werden.

Details für die kuratierten TIER 1 LLM finden sich in den Nutzungsbedingungen der jeweiligen Anbieter (z.B. Mistral oder OpenAI).

5. Die Verarbeitung personenbezogener Daten im Sinne der [DSGVO] sowie sonstiger sensibler Daten bedarf einer Rechtmäßigkeitsprüfung im Hinblick auf geltende Gesetze und interne Datenschutzregelungen. Diese muss in Abstimmung mit der Datenschutzkoordination und dem Datenschutzbeauftragten des FVB erfolgen.

6. Wenn personenbezogene Daten in Forschungsprojekten verarbeitet werden dürfen, sind diese je nach Art und Beschaffenheit der Daten in geeigneter Weise zu anonymisieren oder pseudonymisieren. Darüberhinaus sind die im Ergebnis des Genehmigungsprozeßes auferlegten Pflichten, etwa hinsichtlich Datensparsamkeit, Datensicherheit, Dokumentation, Löschfristen und Betroffenenrechte, umzusetzen.

7. Datenschutzvorfälle: Sofern es zu einer Verletzung des Schutzes personenbezogener Daten gekommen ist, ist diese unverzüglich der Datenschutzkoordination zu melden und sicherzustellen, dass der Schutz wieder hergestellt wird.

8. Bei Unklarheiten oder konkreten Fragestellungen stehen die Datenschutzkoordination sowie der Datenschutzbeauftragte unterstützend zur Verfügung.

TIER 2: Abhängig vom Nutzungsfall, haben die Nutzer von TIER 2 Projekten eine besondere Sorgfaltspflicht (z.B. Kuratierung genutzter LLM-Modelle, Übersicht geteilter/generierter Daten, Konfiguration der LLM-Services (opt-out Training, Datenspeicherung)). Bei Nutzung externer Anbieter in Drittländern (außerhalb von EU) werden die Vorgaben der [DSGVO] zu Drittlandübermittlungen durch Datenschutzkoordination geprüft.

Softwareentwicklung und Urheberrecht

9. GenKI kann zur Unterstützung bei der Entwicklung von Software (z. B. Code, Shell-Skripte, Webseiten) eingesetzt werden. Die Verantwortung für *Funktionalität, Korrektheit, Sicherheit, Urheberrecht und Lizenzkonformität* der resultierenden Software liegt bei den Entwickelnden/KI-Nutzenden.

10. Es ist darauf zu achten, dass in TIER 1/2 keine vertraulichen Quelltexte oder proprietären Algorithmen offengelegt werden, sofern dies nicht ausdrücklich freigegeben ist. Bei Open-Source-Projekten sind die jeweiligen Lizenzbestimmungen zu beachten. Hinweis: *GenKI kann urheberrechtlich geschützte Inhalte beinhalten/reproduzieren. Viele kommerzielle und offene LLM besitzen keine transparente Dokumentation der verwendeten Trainingsdaten, was die Prüfung von Urheberschutzfragen deutlich erschwert.*

TIER 2: Der Einsatz von KI bei der Softwareentwicklung soll nachvollziehbar dokumentiert werden, etwa durch aussagekräftige Commit-Texte. Nutzende prüfen dabei die Vereinbarkeit der verwendeten Komponenten mit Lizenz- und Urheberrechtsvorgaben. Urheberrecht entsteht ausschließlich durch menschliche schöpferische Leistung; rein KI-generierte Codes sind nicht urheberrechtlich geschützt. Für agentische Systeme, die automatisch Code generieren, ausführen oder externe Tools ansteuern, sind zusätzliche technische Schutzmaßnahmen empfohlen (z. B. Sandbox-Umgebungen, Einschränkung kritischer Aktionen, Logging).

Lehre und Publikationen

11. In der Lehre kann GenKI sowohl von Lehrenden als auch von Lernenden unterstützend eingesetzt werden. Prüfungs- und Studienordnungen sowie die jeweiligen Bestimmungen der Universitäten sind zu beachten. Täuschungshandlungen (z. B. Einreichen von KI-generierten Texten als eigene Leistung ohne Kennzeichnung) sind unzulässig.

12. Für wissenschaftliche Publikationen ist die Nutzung von GenKI transparent zu machen, sofern dies von Zeitschriften, Verlagen oder Förderorganisationen gefordert wird. Die jeweils gültigen Richtlinien (z. B. zum Einsatz von KI in Manuskripten) sind zu beachten.

TIER 2: Die Nutzung in Publikationen (z. B. umfangreiche KI-gestützte Datenanalysen oder komplexe agentische Workflows) ist sicherzustellen, dass die Ergebnisse nachvollziehbar, reproduzierbar und unabhängig prüfbar bleiben. Dies umfasst insbesondere die Dokumentation der genutzten Modelle, Konfigurationen und Datenverarbeitungsschritte.

Schlussbemerkungen

Die vorliegende Richtlinie soll regelmäßig überprüft und bei Bedarf angepasst werden. Insbesondere die europäische und nationale Regulierung von KI, neue Angebote externer Anbieter sowie technische Entwicklungen (z. B. neue Modellklassen, agentische Systeme) erfordern eine fortlaufende Weiterentwicklung.

Im Rahmen der Entwicklung eigener KI-Modelle und KI-Systeme sind, soweit diese nicht ausschließlich zu Forschungszwecken im Sinne von Art. 2 Abs. 6 KI-VO eingesetzt werden, ergänzend die Anbieter-Pflichten der KI-VO zu prüfen und zu beachten. Änderungen am TIER-System, an den zugelassenen Modellen oder an den Rollen und Zuständigkeiten (z. B. Ombudsperson, Datenschutzkoordination, Beirat IT) werden in den entsprechenden Richtlinie TIER 1/TIER 2 sowie in begleitenden Prozessdokumenten festgehalten. Diese Richtlinie verweist in Zweifelsfällen auf diese ergänzenden Dokumente.

Für Fragen zur Auslegung dieser Richtlinie, zu konkreten Anwendungsfällen oder zu geplanten TIER 2-Projekten stehen insbesondere Ombudsperson, Datenschutzkoordination, Forschungssoftwareingenieur (RSE) und IT-Abteilung als Ansprechpersonen zur Verfügung.

Autoren: Christian Merdon[2], Dirk Peschka[3], Laura Prieto Saavedra[4]

Literatur

[EU AI Act] EU KI-Gesetz / Artificial Intelligence Act (Verordnung (EU) 2024/1689). https://artificialintelligenceact.eu/de/ und https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[DSGVO] *Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr (Datenschutz-Grundverordnung)*. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[GWP] Deutsche Forschungsgemeinschaft (DFG): *Kodex „Leitlinien zur Sicherung guter wissenschaftlicher Praxis“*. https://www.dfg.de/gwp. Forschungsverbund Berlin e.V. (FVB): *„Verfahrensordnung bei Verdacht auf wissenschaftliches Fehlverhalten“* https://www.fv-berlin.de/ueber-uns/hinweise-kritik/gute-wissenschaftliche-praxis

[EU Guideline] European Commission: *Living Guidelines on the Responsible Use of Generative AI in Research*. https://research-and-innovation.ec.europa.eu/document/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en.

[Leibniz Guideline] Leibniz-Gemeinschaft: *„Empfehlung zur Sicherung der guten wissenschaftlichen Praxis beim Umgang mit Künstlicher Intelligenz“*. DOI: https://dx.doi.org/10.5281/zenodo.14420893.

[2] Datenschutzkoordinator des WIAS
[3] Ombudsperson des WIAS
[4] Research Software Engineer (RSE) des WIAS