ZE Zentrum
VE verantwortungsbewusste
DI Digitalisierung

Centre Responsible Digitality

Zur forschungsethischen Begutachtung von KI-Forschungsprojekten

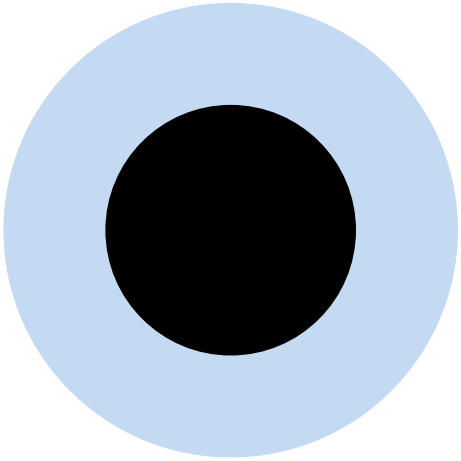
Handreichung zur Unterstützung der Arbeit von Ethikkommissionen an Hochschulen

Version 1.0

Oktober 2022

ZE Zentrum
VE verantwortungsbewusste
DI Digitalisierung

Centre Responsible Digitality



Sie finden die Handreichung auch zum Download.
https://zevedi.de/aktivitaeten/dokumente

Forschungsethik

Die forschungsethische Begutachtung von Forschungsprojekten, die Versuchspersonen einbeziehen, personenbezogene oder anderswie kritische Daten generieren/kaufen/verwenden oder aber eine bedenkliche Form der Nutzung der Forschungsergebnisse erwarten lassen („Dual Use“), ist in der Wissenschaft üblich. Vielfach schreiben Forschungsförderer oder Publikationsorgane Ethikbegutachtungen sogar explizit vor. Forschungseinrichtungen richten zu diesem Zweck Ethikkommissionen ein und definieren geeignete Begutachtungsverfahren. Die Begutachtung ist im Grundsatz freiwillig. Ein Ethikvotum einzuholen, gehört dennoch in typischen Fällen zum methodischen Standard. Zu beachten ist, dass Ethikvoten kein Ersatz sind für die individuelle Verantwortung des oder der Einzelnen für die Güte und auch die ethikrelevanten Aspekte ihrer konkreten wissenschaftlichen Arbeit.

Forschungsethik kann als Form der Qualitätssicherung betrachtet werden. Sie umfasst aber auch Aspekte guter wissenschaftlicher Praxis, professionsethische Standards sowie das geltende Recht.[1] Ebenso kann sie einrichtungs- oder hochschuleigene Bestimmungen (z.B. ein explizites Leitbild o.ä. oder eine Zivilklausel) sowie Vereinbarungen mit (und damit legitime Erwartungen seitens) Kooperationspartnern oder normative Vorgaben eines Forschungsförderers (z.B. dem sog. Gender Data Gap entgegenzuwirken oder Tierversuche zu unterlassen) mit einbeziehen.

„KI“ als Begutachtungsgegenstand

Forschungen im Bereich der sogenannten Künstlichen Intelligenz (KI oder AI) sind nicht völlig neu. Sie haben in den letzten Jahren aber eine veränderte Eingriffstiefe gewonnen. Dies liegt an der Verbindung von einerseits anspruchsvoller Algorithmik und andererseits sehr großen, oft ‚alltagsnah‘ gewonnenen Datenmengen. KI-Szenarien (u.a. sog. maschinelles Lernen, „tiefes“ Lernen) erbringen – in der Forschung zunächst modellhaft – adaptive, bis zu einem gewissen Grad „autonom“ gefundene Lösungen. Diese können wiederum in eine Vielzahl von Diensten und Produkten einfließen. Die potenzielle Reichweite von Technikfolgen bzw. ungewollten Techniknebenfolgen (etwa soziale Diskriminierungseffekte) oder auch von Qualitätsmängeln von KI-Lösungen ist daher in einigen Feldern, zum Beispiel im Medizin-, Finanz- oder Sicherheitsbereich ungewöhnlich groß.[2]

Forschungsethik hat hier Nachholbedarf, zumal forschungsethische Standards im für KI-Lösungen ausschlaggebenden Fach Informatik generell noch nicht tief verankert sind. Es gibt hier kaum „Professionskultur“, an die angeknüpft werden kann. Auch die Gesellschaft ist auf die Fähigkeiten von KI-Technologie noch nicht eingestellt. Gesetzliche Sicherheits- und Haftungsbestimmungen für Produkte, die KI enthalten, müssen derzeit erst entstehen.

Viele Fragen, die KI-Forschungsprojekte aufwerfen, lassen sich mittels „klassischer“ Forschungsethik beantworten. Einige sind jedoch spezifischer Art.[3] Zu solchen KI-spezifischen Problemen versucht das nachfolgende Papier Bewertungskriterien – insbesondere für Forschungsvorhaben aus dem informatischen Spektrum – zu formulieren. Ziel ist es, Ethikkommissionen in ihrer Arbeit zu unterstützen.[4]

Klassische Prüfkriterien der Ethikbewertung von Forschung

Wie für andere Forschungsprojekte gelten auch für informatische Forschungsvorhaben u.a.

- das Verbot körperlicher oder psychischer Schädigung (sowie die Pflicht zur Abklärung individueller Risiken) von Versuchspersonen [VP]; sind Kranke einbezogen, ist das Votum einer medizinischen Ethikkommission erforderlich (hier gelten die Standards der BÄK/ZEKO[5]), das Vorgehen bei Arzneimittelprüfung ist im Arzneimittelgesetz geregelt;
- die gesetzlichen Bestimmungen zum Umgang mit Gefahrstoffen und zu Hochsicherheitslaboren;
- die üblichen Regeln (z.B. der DGP) für die Aufklärung und Einwilligung von Versuchspersonen (einschließlich der nachträglichen Aufklärung von Täuschungsversuchen);[6]
- die fachübergreifenden Standards bezüglich Repräsentativität von Stichproben sowie ggf. zu vermeidender Diskriminierungseffekte bzw. Benachteiligung vulnerabler Gruppen;
- die zu Zwecken der Reproduzierbarkeit bzw. Nachvollziehbarkeit von Forschungsergebnissen erforderlichen, zur gewählten Forschungsmethode/Forschungsform passenden Dokumentations- und Archivierungspflichten;
- die Bestimmungen der DSGVO zu Datenschutz, Datensparsamkeit und Zweckbindung der Erhebung bzw. Nutzung von personenbezogenen Daten einschließlich des Gebotes, diese nach Gebrauch zu löschen;
- die Standards guter wissenschaftlicher Praxis bezüglich methodischer Kompetenzen der Projektleitung, Autorschaft, akademischen Abhängigkeitsverhältnissen;[7]
- die Standards zur Transparenz von Industriekooperationen, zur Veröffentlichung auch der in Kooperationen mit der Industrie auf wissenschaftlicher Seite gewonnenen Ergebnisse etc.;
- das Gebot, die Reputation von öffentlicher Wissenschaft nicht zu Zwecken der kommerziellen Werbung zu missbrauchen;

- die Regeln und Standards eines zeitgemäßen Forschungsdatenmanagements und der Qualitätssicherung von Daten im gesamten Forschungszyklus (auch unter dem Gesichtspunkt der digitalen Souveränität/Datensouveränität der Wissenschaft);
- das Gebot, von überflüssiger bzw. absehbar nutzloser Forschung abzusehen;
- die gesetzlichen Bestimmungen des StGB und des Außenwirtschaftsgesetzes.

Ebenso muss ein Forschungsprojekt, um ein Ethikvotum zu erhalten, vollständig und u.a. hinsichtlich Methoden, Arbeitsteilung im Team sowie Verantwortlichkeiten und ggf. Haftungsfragen hinreichend klar beschrieben sein.

Prüfkriterien einer KI-spezifischen Ethikbewertung von Forschung

Es ist zu empfehlen, die folgenden KI-spezifischen forschungsethischen Fragen zu prüfen bzw. zu betrachten, sie entstehen in fünf Hinsichten:

1. KI als Gegenstand des Forschungsprozesses selbst

1.1. Dokumentation. Wie werden Daten, Algorithmen und das dynamische Zusammenspiel von beidem dokumentiert?[8] Mindestens erforderlich sind die Charakterisierung der Daten (Art, Umfang, Herkunft), die Dokumentation der Software und Trainingsmethoden, die Spezifikation der Hardware, die Dokumentation relevanter Guidelines etc. sowie im Fall sog. „Hochrisikosysteme“[9] auch die automatisch erstellten Dokumentationen des Rechenvorgangs.
Sind die Dokumentationen nicht hinreichend, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

1.2. Reproduzierbarkeit. Welche inhärenten Grenzen der Nachvollziehbarkeit, Reproduzierbarkeit und der Prognostizierbarkeit des „Verhaltens“ der KI-Lösung sind gegeben – und wie trägt der Versuchsaufbau bzw. wie trägt das Projekt den damit einhergehenden Unschärfen/Qualitätsmängeln Rechnung? Ein Ethik-Antrag muss auf diese Fragen aussagekräftige Antworten geben.
Sieht das Forschungsdesign keine Teststrategien für die Reproduzierbarkeit des Lösungsverhaltens vor (ggf. mit Implementierung einschlägiger Tests), ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

1.3. Nachnutzbarkeit, Forschungsdatenmanagement. Wie werden die Teile einer KI-Lösung zur Nachnutzung archiviert? Mindestens erforderlich ist eine Archivierung einschließlich aller Dokumentationen, KI-spezifischer Metadaten (u.a. Risikoklassen, EU-Konformität) sowie von Ethikanträgen und -voten. Neben der Einhaltung der FAIR-Standards[10] („Findable, Accessible, Interoperable, Reusable“) müssen Metadaten auch die Provenienz der wesentlichen Komponenten einer KI-Lösung bzw. eines KI-Experiments enthalten.[11]
Sind die erforderlichen Metadaten nicht erstellt sowie deren Archivierung nicht vorgesehen, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

1.4. Datenschutz. Wie wird bei der Verwendung von anonymisierten personenbezogenen Daten die direkte oder indirekte (Re-)Identifizierung verhindert? Wie wird die Einhaltung der DSGVO bei der Verwendung von personenbezogenen Daten gewährleistet? Wurde bei der Erhebung von personenbezogenen Daten in die Verwendung eingewilligt oder besteht ein überwiegendes öffentliches Interesse an der Datennutzung? Ist die Verarbeitung personenbezogener Daten aus wissenschaftlichen Gründen erforderlich? Wird dabei der Grundsatz der Datenminimierung (Art. 5 DSGVO) beachtet und sind ausreichende Sicherungsmaßnahmen bei der Archivierung getroffen (IT-Sicherheit, Präventivmaßnahmen gegen Sabotage, Datendiebstahl etc.)? Ein Ethik-Antrag muss diese Problemstellungen schlüssig lösen.[12]

Fehlt eine differenzierte Erläuterung zur Konformität mit der DSGVO, Art. 5 – 23 (einschließlich des Ausschlusses von model inversion) im Blick auch auf Anhang IV der EU-KI-Verordnung und den einschlägigen, von verschiedenen Seiten empfohlenen Systemdatenschutz, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

1.5. Wirksamkeit der Einwilligung. Sind, wo VP oder Datengeber eine Datennutzung autorisieren, die Aufklärungsbögen insgesamt sowie in der Darstellung der Typik und der Ziele der verwendeten KI-Lösung selbst hinreichend verständlich formuliert? Enthält der Aufklärungsbogen hinreichend Information über den Verarbeitungszweck, die Weiterverwendung und den Zeitpunkt der Datenlöschung?
Fehlt eine hinreichend (standardgemäße) verständliche Erläuterung und Nachvollziehbarkeit als Voraussetzung einer wirksamen Autorisierung, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

1.6. Datenbeschaffenheit. Welche Typik und welche Qualität haben die verwendeten Daten? Eignen sich die Daten zum „Training" von Algorithmen? Wie wirken sich die Datenbeschaffenheit einerseits, Daten in der Anwendung andererseits bzw. hierzu ggf. fehlendes Wissen auf die Erträge der angewendeten Methoden und letztlich die Qualität der Forschungen/Forschungsergebnisse aus?[13]
Wird keine Klarheit über die Beschaffenheit von Daten, insbesondere von Trainingsdaten, hergestellt, deren Beschaffenheit sich aber (u.a. als Ursache unbemerkter Biases) auf die Qualität der Forschungsergebnisse auswirken, bzw. stehen Vorsorgemaßnahmen zur Vermeidung von Biases aus, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

1.7. Datenkauf/„Lieferketten". Ist die Provenienz gekaufter Datensätze geklärt und sind diese Datensätze unter forschungsethisch vertretbaren Bedingungen (nämlich mindestens den europäischen, idealerweise den in Deutschland geltenden Standards) gewonnen worden? Hierzu sind ggf. Nachweise erforderlich.

Fehlen (hinsichtlich der Einhaltung europäischer und nationaler Sicherheitsstandards sowie zu forschungsethisch gerechtfertigten Bedingungen ihrer Gewinnung aussagekräftige) Provenienznachweise (oder Zertifikate) zu akquirierten Datensätzen, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

1.8. Algorithmen. Welche Typik haben die verwendeten Algorithmen? Bringt der verwendete Typ von Algorithmen typische Risiken mit sich, insbesondere im gewählten methodischen Kontext und mit Blick auf die Datendomäne (z.B. hohe Rate nicht nur von falsch negativen, sondern auch von falsch positiven Ergebnissen, leichte Irritierbarkeit durch überraschende Daten)?
Fehlt eine überzeugende Typisierung der verwendeten Algorithmen hinsichtlich ihrer Stärken und Schwächen, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

1.9. Diskriminierungsfreiheit. Können die Beschaffenheit von Trainingsdaten und/oder Algorithmik im Ergebnis zu Diskriminierungseffekten führen oder eine existierende Diskriminierung (von Individuen, sozialen Gruppen etc.) verstärken?[14] Spiegeln die Daten das für die Forschungsfrage relevante tatsächliche Umfeld repräsentativ wieder?
Fehlen auf Daten- oder Modellebene Vorsorgemaßnahmen gegen eine Diskriminierung oder eine Verstärkung bestehender Diskriminierung vulnerabler Gruppen, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

1.10. Nachhaltigkeit. Werden die Mittel der öffentlichen Forschungsförderung effizient eingesetzt? Wie teuer ist der geplante Rechnereinsatz/die Rechenzeit? Sind hier ökonomische bzw. ökologische Effektivitätsgewinne möglich?[15]
Werden im Falle eines aufwendigen Rechnereinsatzes die eingesetzten Ressourcen nicht bilanziert oder erscheint dieser Ressourceneinsatz suboptimal, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

2. Verwendung von proprietären KI-„Tools“ im Forschungsprozess

2.1. Proprietäre Toolboxes, „Blackboxing“. Kommen im Rahmen eines Forschungsprojekts proprietäre KI-Komponenten/KI-„Tools“ zum Einsatz und in welchem Ausmaß entstehen hierdurch nicht behebbare Intransparenzen für den Forschungsprozess (sog. Blackboxing Probleme)? Im Ethik-Antrag muss beschrieben werden, welches Gewicht den Anteilen des Forschungsworkflows zukommt, die der Kenntnis der Forschenden und ihrem Zugriff entzogen sind, warum durch die Wissenschaft selbst qualitätsgesicherte Alternativen nicht genutzt werden können und wie die Forschenden eine Validität möglicher Ergebnisse trotz des Einsatzes proprietärer KI-Werkzeuge sicherstellen.
Fehlen die Dokumentation von Intransparenzen proprietärer Toolboxes sowie Aussagen dazu, wie trotz Blackbox-Effekten die Validität der Ergebnisse gewährleistet werden kann, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

2.2. Datenschutz und Datenintegrität auf Seiten der Anbieter von KI-Werkzeugen für die Wissenschaft. Können die Anbieter kommerzieller oder aus anderen Gründen auf proprietären Daten/Software basierenden Analysediensten die Einhaltung europäischen Rechts und europaweiter Standards garantieren? Auch die Anbieter von Forschungswerkzeugen müssen die DSGVO einhalten und die Datenverarbeitung innerhalb der EU zusichern. Maßgeblich hierfür sind Verträge, nicht allein Angaben im „www“.
Wird die Einhaltung europäischer Standards seitens der Anbieter von KI-Forschungswerkzeugen/Analysetools nicht in juristisch belastbarer Form dokumentiert, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

3. KI-Lösungen in einer überwiegend durch nichtinformatische Forschungsfragen geprägten Domäne

3.1. Sektorenspezifische Forschung und Anwendung. In welcher wissenschaftlichen Domäne[16] bzw. in welchem gesellschaftlichen „Sektor“ (z.B. Medizin, Mobilität, Zahlungsverkehr, schulische Bildung, Sicherheit) sollen die Forschungsergebnisse zum Einsatz kommen und wie wirkt sich die Typik der Anwendungsdomäne auf die Risikobilanz der geplanten Forschungen bzw. die zu erwartende Risikoklasse möglicher Anwendungen[17] aus? Entstehen nicht abgesehene Risiken beim Transfer der Forschungsergebnisse/Lösungen in andere Domänen bzw. Sektoren? Besteht die Notwendigkeit, den Einsatz der erforschten/entwickelten KI-Lösung auf bestimmte Domänen bzw. Sektoren als Anwendungsfeld zu begrenzen? Lässt sich dies bereits *by design* implementieren?
Ohne Angabe zu Sektoren, in welchen die Forschungsergebnisse zum Einsatz kommen, und Angaben zu möglichen Risiken eines Transfers auf andere Sektoren sind Risikobilanzen im Bereich anwendungsnaher KI-Forschung unvollständig. In solchen Fällen ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

3.2. Domänenspezifische Forschungskompetenz. Domänenspezifische Forschungsfragen machen ggf. domänenspezifische KI-Lösungen erforderlich. Besitzt das Forschungsteam in den dann gefragten Feldern hinreichendes Domänenwissen bzw. auf die Domäne zugeschnittene Methodenkompetenz? Wo Forschende, die das Projekt durchführen (z.B. Informatiker*innen, Datenwissenschaftler*innen) eine zu geringe Kenntnis der Besonderheiten der (Daten-)Domänen bzw. „Use Cases“, zu welchen geforscht wird, erwarten lassen, ist eine interdisziplinäre Beteiligung von Wissenschaftler*innen mit Expertise für die analysierte Domäne (z.B. Sozialwissenschaftler*innen, Wirtschaftswissenschaftler*innen, Mediziner*innen) geboten.
Fehlt angesichts domänenspezifischer Problemlagen (z.B. in *use cases*) die erforderliche zusätzliche Fachkompetenz anderer Disziplinen

(Fachwissen, Methodenkompetenz), ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

3.3. Interdisziplinäre Teams und interdisziplinäre Publikation. KI-Forschung erlebt eine schnelle Evolution. Vor diesem Hintergrund ist sicherzustellen, dass Forschungsergebnisse zu überwiegend durch nichtinformatische Forschungen geprägte Domänen so publiziert werden, dass sie auch in der Forschungskultur der Domäne (ggf. kritisch) rezipiert werden können.[18] Interdisziplinäre KI-Forschungsteams sollten in Ethik-Anträgen Aussagen zur Publikationsstrategie treffen. Breit sichtbares (informatisch und in den Domänen rezipiertes) Publizieren ist exklusiven Publikationen in sehr kleinen (Teil-)Communities vorzuziehen.
Ist bei KI-Forschung mit erwartbar hohem Impact für nichtinformatische Domänen eine nur auf die KI-interne Fachkommunikation beschränkte Publikationsstrategie vorgesehen, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

3.4. Technikfolgen, „Gesellschaft“. Abschätzungen eventueller sozialer Folgen des Einsatzes einer KI-Lösung erfordern in der Regel nicht nur individualpsychologische, sondern sozialwissenschaftliche Expertise. Bei Fragestellungen, die nicht die Schnittstelle zu individuellen Nutzenden („Nutzungsforschung“), sondern gesellschaftliche Fragen betreffen, ist daher die Beteiligung von Sozialwissenschaftler*innen vorzusehen.
Wird eine Bewertung gesellschaftlicher Folgen des Einsatzes von KI-Systemen vorgesehen ohne sozialwissenschaftliche oder normative Kompetenz einzubinden, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

4. Zu erwartende Anforderungen an künftige Produktmerkmale (z.B. „EU-Konformität“) in der anwendungsnahen KI-Forschung

4.1. Risikoklassen/Hochrisikosysteme. In Produkten welcher Risikoklasse und welcher Einsatzbereiche werden die Forschungsergebnisse absehbar zum Einsatz kommen? Der Bezug zum Risikoklassensystem der spezifischen Bereiche im *Anhang III* der in Entwurfsfassung vorliegenden EU-KI-Verordnung[19] sollte ggf. bereits im Forschungsprozess hergestellt werden, weil hier unter Umständen zusätzliche Anforderungen an den Forschungsprozess entstehen (z.B. [Stufe 3 und 4] die Dokumentation der Namen aller am Forschungsprozess beteiligten Personen).
Fehlt bei der Planung/Festlegung von zu erwartenden Produktmerkmalen („Pflichtenheft“) eine explizite Bezugnahme auf die EU-Risikoklassifizierung (App. III), ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

4.2. EU-Konformität. Die EU-Zulassung von KI-Systemen setzt auch jenseits der Risikoklassen EU-Konformität voraus. So gehört Kontrollierbarkeit zu den gesetzlichen Anforderungen an künftige KI-Produkte. Dies schließt Risikomanagement (Schadenshöhe, Schadenswahrscheinlichkeit, Disponibilität von Risiken) und die Frage nach Grundrechtsverletzungen (z.B. Diskriminierungen) ein.
Es kann die ethische Vertretbarkeit von Forschung schmälern, wenn eine anwendungsnahe Entwicklung absehbar nicht EU-konform ausgelegt werden und damit in der EU keine Produktreife erlangen kann.[20]

4.3. Resilienz. Schließen die Forschungen die Frage einer in der Praxis (zu) leichten Irritierbarkeit der KI-Lösung in den Produkten mit ein, in welchen sie künftig zum Einsatz kommen? Ist geklärt, wie erheblich die Folgen ggf. wären?
Bezieht das Forschungsdesign Planungen hinsichtlich der Resilienz zu entwickelnder Systeme mit ein? Wo dieser Aspekt bei anwendungs-

naher, auf konkrete Produkte hinarbeitender Forschung fehlt, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

4.4. Erklärbarkeit. Ist das für die Produktentwicklung/eine künftige Zulassung erforderliche Maß an „Erklärbarkeit“ (Explainabilitiy) des Forschungsvorhabens wie auch seiner Ziele und insbesondere der Ziele der verwendeten KI-Typik sichergestellt?
Ist eine solche Erklärbarkeit nach Anhang IV der EU-KI-Verordnung in Verbindung mit einschlägig delegierten Zertifizierungsanforderungen (z.B. VDE SPEC[21]) nicht sichergestellt, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

4.5. Digitale Souveränität/Option der Nichtnutzung. Tragen die KI-Lösungen, an denen geforscht wird, am Markt perspektivisch (ggf. indirekt) zu Lock-In Effekten und anderen Formen einer Abhängigkeit von bestimmten Technologien bzw. einer Unvermeidlichkeit von Überwachung bei, die im Alltag nicht mehr umgangen werden können? Schließen Entwicklungspfade die Nicht-Nutzung entstehender neuer Technologien aus?[22]
Wird durch einen Entwicklungspfad, den die Forschung einschlägt, die Option einer Nicht-Nutzung von Systemen eingeschränkt oder verunmöglicht („Alternativlosigkeit“ der Nutzung), ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

5. KI und Dual Use-Konstellationen

5.1. Dual Use. Welche „Dual Use“-Probleme sind mit den geplanten Forschungen sowie den auf ihrer Basis in der Anwendung (weiter-)entwickelten KI-Lösungen zu erwarten? Dual Use-Konstellationen stellen sich in der KI-Forschung schon deshalb schnell ein, weil weltweit Regierungen derzeit in KI-basierte Rüstungs- und Sicherheitstechnologien investieren. Die Prüfung von Dual Use-Optionen der Ergebnisse eines Forschungsvorhabens betrifft daher nicht nur der Kriegsführung dienende Forschung und militärischen Szenarien benachbarte Pfade der Forschungsförderung (etwa Katastrophenabwehr, Kriminalitätsbekämpfung), sondern ggf. auch Forschungen für andere Domänen.[23]

Die Forschung in Deutschland hat hier die Wahrscheinlichkeit eines Einsatzes von Forschungsergebnissen zu unfriedlichen und auf andere Weise verfassungswidrigen Zielen zu minimieren. Zu fragen ist daher: Wie werden unerwünschte (z.B. unfriedliche, Grundrechte gefährdende, Demokratie und Marktstabilität gefährdende oder ökologisch gefährliche) Nutzungen nach Möglichkeit ausgeschlossen oder aber unwahrscheinlicher oder unattraktiver gemacht?
Werden Dual Use-Konstellationen nicht angemessen reflektiert, ist ein Nachbesserungen forderndes oder negatives Ethikvotum angezeigt.

5.2. Zivilklauseln. Hat sich eine Einrichtung eine Zivilklausel gegeben, kann die Vereinbarkeit geplanter Forschungen mit dem Wortlaut dieser Zivilklausel Gegenstand der Ethikbegutachtung sein.[24]
Ein festgestellter Zivilklauselverstoß macht ein Nachbesserungen forderndes oder negatives Ethikvotum nötig.

Weitere Hinweise für Antragstellende

„Grundlagenforschung“? Forschungen an KI-Verfahren sind derzeit insbesondere im Bereich der Algorithmik oft (noch) „generischer“ Natur, es werden Grundlagen erforscht. Durch das stets notwendige Zusammenspiel mit konkreten Daten oder Use Cases ist Forschungsethik jedoch auch im Stadium sogenannter Grundlagenforschung relevant.

Zeitplanung. Eine forschungsethische Evaluation von Projekten/Szenarien im Feld „KI“ kann komplex sein. Ethikkommissionen sollten daher frühzeitig um ein Votum gebeten werden. Zu einer guten Forschungsplanung gehört es, die entsprechende Vorlaufzeit hierfür vorzusehen.

Zuständigkeiten. Zuständig sind grundsätzlich die Ethikkommissionen der Forschungseinrichtungen, an welchen die Forschenden tätig sind. Für medizinische Forschungsvorhaben ist das Votum einer medizinischen Ethikkommission erforderlich. Insbesondere in der Verbundforschung müssen nichtmedizinische Anteile eines Forschungsvorhabens

eventuell gesondert begutachtet werden. Ein Verzeichnis der nichtmedizinischen Ethikkommissionen („KEF“) pflegt die Leopoldina.

Beratung. Beratungskapazitäten zur KI-Forschungsethik sind in den Einrichtungen des Wissenschaftssystems derzeit erst im Aufbau. Forschenden wird empfohlen, sich an die Ethikkommissionen ihrer jeweiligen Einrichtung zu wenden. Vorsitzenden, Mitgliedern und Geschäftsstellenmitarbeiter*innen von Ethikkommissionen bietet das Zentrum verantwortungsbewusste Digitalisierung (ZEVEDI) [office[at]zevedi.de] die Möglichkeit zur Vernetzung an.

Oktober 2022,
ZEVEDI-Projektgruppe *Normordnung künstlicher Intelligenz* (NOKI)

Anmerkungen

[1] Bei grundsätzlichem Interesse am Thema sei zur Abrundung des Bildes einer auf Probleme der Digitalität sich allerdings spät einstellenden Ethik-Praxis empfohlen: Arbeitskreis Medizinischer Ethik-Kommissionen in der Bundesrepublik Deutschland e.V., Journal of Academic Ethics, Jahrbuch Wissenschaft und Ethik.

[2] Vgl. den Bericht über die Auswirkungen künstlicher Intelligenz, des Internets der Dinge und der Robotik in Hinblick auf Sicherheit und Haftung der Europäischen Kommission, Brüssel 19.02.2020.

[3] Eine Einordnung von KI für die wissenschaftliche Forschung erbringen Gethmann et al.: Künstliche Intelligenz in der Forschung. Neue Möglichkeiten und Herausforderungen für die Wissenschaft, Berlin 2022.

[4] Die Datenethikkommission formuliert in ihrem Gutachten Handlungsempfehlungen für den Umgang mit Daten und algorithmischen Systemen, die jedoch nicht spezifisch für die wissenschaftliche Forschung sind. Datenethikkommission der Bundesregierung: Gutachten der Datenethikkommission, Berlin 2019.

[5] Zentrale Kommission zur Wahrung ethischer Grundsätze in der Medizin und ihren Grenzgebieten (Zentrale Ethikkommission, ZEKO) bei der Bundesärztekammer: Stellungnahmen.

[6] Berufsverband Deutscher Psychologinnen und Psychologen e.V. / Deutsche Gesellschaft für Psychologie e.V.: Berufsethische Richtlinien, Berlin 2016.

[7] Deutsche Forschungsgemeinschaft: Leitlinien zur Sicherung guter wissenschaftlicher Praxis. Kodex, Bonn 2019.

[8] Generelle Anforderungen an Dokumentation als Teil guter wissenschaftlicher Praxis werden in Leitlinie 12 des DFG-Kodex benannt (a.a.O.).

[9] Wie zuvor schon diverse KI-Ethik-Expertisen teilt auch KI-Verordnung der Europäischen Union „KI-Systeme“ in Risikoklassen ein. Der risikobasierte Ansatz unterscheidet Anwendungen von KI in solche die i) ein unannehmbares, ii) ein hohes oder iii) ein geringes bzw. minimales Risiko darstellen. Vgl.: Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, Brüssel, 21.04.2021, S. 15. Zudem werden in Art. 5 Abs. 1 bestimmte Praktiken im Bereich der künstlichen Intelligenz verboten wie etwa

Techniken der unterschwelligen Beeinflussung, das Ausnutzen von Schwächen bestimmter Personengruppen, die Bewertung oder Klassifizierung der Vertrauenswürdigkeit natürlicher Personen durch Behörden oder biometrische Echtzeit-Fernidentifizierungssysteme in öffentlich zugänglichen Räumen zu Strafverfolgungszwecken. Art. 6 verweist auf die Anhänge II und III für die Klassifizierung als Hochrisiko-KI-System. Anhang II der EU-KI-Verordnung listet die Harmonisierungsrechtsvorschriften für Produkte auf, unter denen KI-Systeme als hochriskant klassifiziert werden. Anhang III der EU-KI-Verordnung unterscheidet für die Einordnung von Hochrisiko-KI-Systemen etwa folgende Bereiche: 1) Biometrische Identifizierung und Kategorisierung natürlicher Personen, 2) Verwaltung und Betrieb kritischer Infrastrukturen, 3) Allgemeine und berufliche Bildung, 4) Beschäftigung, Personalmanagement und Zugang zur Selbstständigkeit, 5) Zugänglichkeit und Inanspruchnahme grundlegender privater und öffentlicher Dienste und Leistungen, 6) Strafverfolgung, 7) Migration, Asyl und Grenzkontrolle, 8) Rechtspflege und demokratische Prozesse.

Für Hochrisiko-KI-Systeme müssen die Protokollierungsfunktionen nach Art. 12 Abs. 4 zumindest folgende Informationen erfassen: a) Aufzeichnung jedes Zeitraums der Verwendung des Systems, b) die Referenzdatenbank, mit der das System die Eingabedaten abgleicht, c) die Eingabedaten, mit denen die Abfrage zu einer Übereinstimmung geführt hat sowie d) die Identität der an der Überprüfung beteiligten natürlichen Personen. Anhang IV der EU-KI-Verordnung listet präzise die Anforderungen an die technische Dokumentation von Hochrisiko-KI-Systemen auf. Anhänge des Vorschlags für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, Brüssel, 21.04.2021.

[10] Vgl. die Erläuterungen zu Leitlinie 13 des DFG-Kodex zu guter wissenschaftlicher Praxis „Herstellung von öffentlichem Zugang zu Forschungsergebnissen“ (a.a.O.) sowie die detaillierte Beschreibung der Prinzipien der FAIR-Initiative.

[11] Vgl. Leitlinie 5 „Nachnutzen und Reproduzieren“ in dem Positionspapier des Wissenschaftsrats: Zum Wandel in den Wissenschaften durch datenintensive Forschung, Köln 2020, S. 42f.

[12] Vgl. hierzu auch die Stellungnahme des Deutschen Ethikrats: Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung, Berlin 2017.

[13] Vgl. Rat für Informationsinfrastrukturen: Herausforderung Datenqualität. Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, zweite Auflage, Göttingen 2019.

[14] In der Empfehlung der UNESCO zu einer Ethik der künstlichen Intelligenz werden in den Absätzen 28-30 die Prinzipien der Fairness und der Nicht-Diskriminierung in Bezug zu dem Ziel einer gerechten Welt bezüglich Information, Kommunikation, Kultur, Bildung, Forschung und sozioökonomischer sowie politischer Stabilität gesetzt; UNESCO: Recommendation on the Ethics of Artificial Intelligence; Paris, 22.11.2021, S. 9.

[15] In der Empfehlung der UNESCO zu einer Ethik der künstlichen Intelligenz wird in Absatz 31 eine umfassende Kenntnis der Implikationen von KI-Technologien auf die verschiedenen Dimensionen von Nachhaltigkeit gefordert vor dem Hintergrund des Ziels nachhaltiger Gesellschaften; UNESCO: Recommendation on the Ethics of Artificial Intelligence; Paris, 22.11.2021, S. 9.

[16] Die sich aktuell erneuernde EU-Gesetzgebung wie auch die entstehenden EU-Infrastrukturen (z.B. GAIA X) sehen die Ausgestaltung von – auch normativ zu unterscheidenden – Sektoren, Domänen bzw. Datenräumen vor. Vgl. auch Anmerkung 9.

[17] Vgl. Anmerkung 9.

[18] In der Empfehlung der UNESCO zu einer Ethik der künstlichen Intelligenz wird in Abs. 110 die Förderung interdisziplinärer Forschung zu und mithilfe von KI gefordert; UNESCO: Recommendation on the Ethics of Artificial Intelligence; Paris, 22.11.2021, S. 21.

[19] Vgl. Anmerkung 9.

[20] Anhang V der EU-KI-Verordnung listet die Angaben auf, die für eine EU-Konformitätserklärung notwendig zu machen sind.

[21] Vgl. den ersten Draft für den VDE SPEC 90012: VCIO based description of systems for AI trustworthiness characterisation, Offenbach am Main, 24.05.2022.

[22] In der Empfehlung der UNESCO zu einer Ethik der künstlichen Intelligenz wird in Abs. 20 die Möglichkeit einer optionalen Nutzung von KI-Systemen gefordert; UNESCO: Recommendation on the Ethics of Artificial Intelligence; Paris, 22.11.2021, S. 7.

[23] Die Deutsche Forschungsgemeinschaft und die Nationale Akademie der Wissenschaften Leopoldina haben Empfehlungen zum Umgang mit sicherheitsrelevanter Forschung vor dem Hintergrund der Dual Use-Problematik entwickelt: Wissenschaftsfreiheit und Wissenschaftsverantwortung. Empfehlungen zum Umgang mit sicherheitsrelevanter Forschung, Bonn/Halle (Saale) 2014.

[24] Die Nationale Akademie der Wissenschaften Leopoldina pflegt eine Liste von Kommissionen, die für Ethik sicherheitsrelevanter Forschung zuständig sind („KEF“): Ansprechpartner und Kommissionen in Deutschland, die für Ethik sicherheitsrelevanter Forschung zuständig sind. DFG und Leopoldina stellen zudem eine Mustersatzung für Kommissionen für Ethik sicherheitsrelevanter Forschung bereit, in der regelungsbedürftige Sachverhalte aus dem Bereich sicherheitsrelevanter Forschung ausgewiesen werden.

ZEVEDI

Das Forschungs- und Kompetenznetz ZEVEDI bündelt die wissenschaftliche Expertise der hessischen Hochschulen zur Analyse normativer Aspekte des digitalen Wandels und trägt zur Gestaltung dieses Wandels bei.

Das Zentrum konkretisiert Verantwortung als wichtigen Gesichtspunkt von Technologieentwicklung und arbeitet daran, diesen umsetzbar zu machen.

Es erbringt Forschungsleistungen, stärkt den Transfer von Wissen in die Wirtschaft und die Gesellschaft hinein und berät die Politik forschungsbasiert zu den Themen Recht, Ethik und Innovation – für eine demokratische und humane Ausrichtung des digitalen Wandels.

ZEVEDI wird gefördert durch die Hessische Ministerin für Digitale Strategie und Entwicklung.

Mitwirkende

Arbeitsgruppe KI-Forschungsethik

Dr. des. Andreas Brenneis
Technische Universität Darmstadt, Redaktion

Prof. Dr. Petra Gehring
Technische Universität Darmstadt, Leitung der AG

Prof. Dr. Christoph Hubig
Technische Universität Darmstadt

Annegret Lamadé
Philipps Universität Marburg

Prof. Dr. Florian Möslein, LL.M. (London)
Philipps Universität Marburg

Dank

Die Arbeitsgruppe KI-Forschungsethik bedankt sich bei folgenden Expertinnen und Experten, die an der Entstehung der Handreichung mitgewirkt haben:

Dr. Christian Geminn Mag. iur.
Universität Kassel

Prof. Dr. Bert Heinrichs
Forschungszentrum Jülich / Rheinische Friedrich-Wilhelms-Universität Bonn

Prof. Dr. Dieter Sturma
Rheinische Friedrich-Wilhelms-Universität Bonn

Impressum

Verabschiedet im Oktober 2022

Zentrum verantwortungsbewusste Digitalisierung (ZEVEDI)
Geschäftsstelle
Technische Universität Darmstadt
Residenzschloss 1
D-64283 Darmstadt

E-Mail office[at]zevedi.de
Web www.zevedi.de

ZITIERVORSCHLAG
Zentrum verantwortungsbewusste Digitalisierung (ZEVEDI): Zur forschungsethischen Begutachtung von KI-Forschungsprojekten. Handreichung zur Unterstützung der Arbeit von Ethikkommissionen an Hochschulen, Darmstadt 2022, 20 S.

ZEVEDI bevorzugt eine gendergerechte Sprache. In Einzelfällen werden Kollektivbezeichnungen gebraucht, die jeweils Personen aller Geschlechter einbeziehen.

Die deutsche Nationalbibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

ISBN 978-3-910468-00-9

ZE
VE
DI

ISBN 978-3-910468-00-9